crete data, particularly methods of categorical data analysis, with emphasis on applications in the social sciences and biomedical sciences. His books include Categorical Data Analysis *(John Wiley, 1990),* An Introduction to Categorical Data Analysis *(John Wiley, 1996), and* Statistical Methods for the Social Sciences *(with B. Finlay, 3rd ed., Prentice Hall, 1997).*

Ivy Liu is a lecturer in the School of Mathematical and Computing Science, Victoria University, Wellington, New Zealand. She received her Ph.D. in statistics from the University of Florida in 1995. Her research interests are in categorical data analysis for sparse data, especially in various extended types of Mantel-Haenszel estimation methods for stratified categorical data.

Event-history analysis of the diffusion of practices in a social system can show how actors are influenced by each other as well as by their own characteristics. The presumption that complete data on the entire population are essential to draw valid inferences about diffusion processes has been a major limitation in empirical analyses and has precluded diffusion studies in large populations. The authors examine the impacts of several forms of incomplete data on estimation of the heterogeneous diffusion model proposed by Strang and Tuma. Left censoring causes bias, but right censoring leads to no noteworthy problems. Extensive time aggregation biases estimates of intrinsic propensities but not cross-case influences. Importantly, random sampling can yield good results on diffusion processes if there are supplementary data on influential cases outside the sample. The capability of obtaining good estimates from sampled diffusion histories should help to advance research on diffusion.

Estimation of Diffusion Processes From Incomplete Data

A Simulation Study

HENRICH R. GREVE
University of Tsukuba

NANCY BRANDON TUMA
Stanford University

DAVID STRANG
Cornell University

1. INTRODUCTION

Much social research examines diffusion processes involving sequential interdependence of behaviors within a population (for reviews, see

AUTHORS' NOTE: *A preliminary version of some of the work reported in this article was presented at the 1993 meetings of the American Sociological Association. Support for this research was provided by the National Science Foundation (SES 8911666, SES 9213152, and SES 9213258) and the Yamada Foundation. We thank Eric Bloch for assisting with the development of the program used to generate pseudorandom diffusion histories and David J. Pasta for his programming assistance with RATE. We also wish to thank the reviewers for their helpful comments on earlier drafts.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 29 No. 4, May 2001 435-467

Rogers 1995; Strang and Soule 1998). Many recent diffusion studies use event-history analysis that includes covariates measuring social similarity or network properties to investigate the paths of social influence in various populations of persons or organizations (e.g., Marsden and Podolny 1990; Strang 1991; Myers 1997; Soule and Zylan 1997). An important limitation of this line of work is the presumed requirement that data are complete, in particular, that the event times and relevant characteristics of actors are measured for *all* members of the relevant population. This requirement has prevented diffusion analysis in many empirical settings of considerable interest.

In this article, we use simulation methods to explore the performance of diffusion estimators when data are incomplete. We examine several sources of incomplete data. First, we study the impact of time aggregation, in which event times are not measured exactly but only with some imprecision. We then investigate the consequences of right and left censoring, in which there are no observations on the timing of either late or early events. Finally, and perhaps most important, we examine the consequences of analyzing data on samples drawn from the population rather than on the entire population. In each instance, our primary aim is to identify patterns of estimation bias and problems of inference to inform empirical research in the short run and to discover which issues require further methodological attention in the longer term.

1.1. MODELING FRAMEWORK AND PRIOR STUDIES

We work with the additive heterogeneous diffusion model proposed by Strang and Tuma (1993). The model is designed to allow analyses of the spread of some behavior (e.g., adoption of an innovation) through a closed population. The members of the population are partitioned into two sets. One set $S(t)$ consists of members of the population who have experienced the event of interest before time t; the second set $\mathcal{N}(t)$ is composed of those who have not yet had the event by time t and so remain at risk. The model specifies the hazard rate for the members of $\mathcal{N}(t)$ as the sum of two components. One component is an individual's hazard rate independent of social influences; it is the ordinary hazard rate found in survival and event-history analyses. The second component comprises the combined social influence of other

actors in the population; in a diffusion process, influence comes from the other actors who have previously had the event. In the diffusion model proposed by Strang and Tuma, these two components are summed, giving

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

Here,

- $\boldsymbol{x}_n$ is a vector of variables describing n's *propensity* to have the event independent of intrapopulation influences;
- $\boldsymbol{v}_n$ is a vector of variables describing n's *susceptibility* to intrapopulation influences;
- $\boldsymbol{w}_s$ is a vector of variables describing the *infectiousness* of s (i.e., the ability of s to influence others in the population); and
- $\boldsymbol{z}_{ns}$ is a vector of variables describing the social *proximity* of n and s.

This model has been employed in a variety of studies of the adoption of social practices (for a review, see Strang and Soule 1998). Among the findings of these studies are that adoption events are more influential when they occur to actors that have network linkages to the potential adopter (Strang and Tuma 1993; Greve 1995, 1996) or are similar to the potential adopter (Strang and Tuma 1993; Davis and Greve 1997; Soule and Zylan 1997; Greve 1998), that heterogeneity in an innovation's value to actors affects susceptibility to social influence (Davis and Greve 1997), and that diffusion of different practices through a given population can depend on different network linkages as well as on different characteristics of the members of the population (Davis and Greve 1997). Thus, the heterogeneous diffusion model is a productive tool for examining the effects of network links on mimetic behavior.

Estimates are obtained by maximizing the logarithm of the likelihood function (Strang and Tuma 1993). Under very general conditions,[1] maximum likelihood (ML) estimators have good properties in large samples; they are asymptotically normal, unbiased, and consistent. Tuma and Hannan (1984, chap. 5) demonstrated that these large sample properties translate well for an exponential model with independent random samples of event histories in the sizes usually available to sociologists (i.e., at least a few hundred cases). Since the model they studied did not include any social influences, it could not be said a priori whether ML would yield high-quality estimates of the parameters in the additive heterogeneous diffusion model in equation (1).

Greve, Strang, and Tuma (1995) performed an extensive Monte Carlo study of the quality of ML estimators of the parameters in (1) when using *complete* data on a population's diffusion history. They explored the effects of different sets of coefficients for covariates, different proximity structures, and model misspecification. They found that ML estimation recovers the parameters of the model from complete data when the model is correctly specified (i.e., includes the actual variables affecting diffusion). They also evaluated ML estimators of the parameters in (1) for various types of specification errors. Including extraneous variables (i.e., ones with no true effect) in the estimated model had very little effect on the ML estimators. As one would expect, the exclusion of a covariate used to generate the diffusion history worsened estimation, especially when the omitted variable affected susceptibility or infectiousness rather than an actor's propensity to experience the event. They also developed strategies for locating effects in the propensity, susceptibility, or infectiousness term of the model.

1.2. INCOMPLETE DATA ON THE POPULATION

The above studies, as well as other research known to us, assume that data on all events within a bounded population are analyzed. The presumption that diffusion analysis requires complete data on the entire population greatly limits the topics susceptible to inquiry because such data are difficult to obtain in many situations and impossible in others. It is thus useful to distinguish the contexts that produce estimation problems from those that do not and to consider what forms of supplementary data can aid estimation. To do so, we investigate the following sources of incomplete data:

1. time-aggregated observations of events,
2. observations that are incomplete on the right (right censored),
3. observations that are incomplete on the left (left truncated and/or left censored),

4. random samples of observations with equal sampling probabilities, and
5. other selected sampling plans.

All of these situations occur often enough to warrant attention. Measurement of the times of events can be imprecise, creating the problem of time aggregation. A diffusion process may not have ended by the time the data are collected, creating the problem of right censoring. Observations on the first part of a diffusion history may be unavailable because data collection begins after the diffusion process started. Diffusion histories for the entire population may be incomplete either intentionally because data were collected for only a sample of the population or unintentionally because information on certain variables is missing for some cases. We denote these conditions as incomplete data (rather than as missing data) since they differ from the ideal situation of complete and correct data but also differ from the classical problem of a data matrix in which some values of variables are missing for some cases.

Sampling is probably the most important condition to study. When actions of members of a population are statistically independent, random sampling drastically reduces the cost of data collection while allowing analysts to make unbiased estimates of parameters characterizing very large populations. Indeed, it is often argued that by collecting data on a sample, investigators can focus on measuring variables accurately and that the greater accuracy of measurements compensates for the loss of precision resulting from the analysis of fewer cases.

Unfortunately, the efficacy of even simple random sampling is in doubt for diffusion studies because of the loss of information on some of the actors who influence members of the sample; omission of right-hand-side data on influential actors can be expected to result in biased parameter estimates. We attempt to identify some simple sampling schemes that allow unbiased estimation of the parameters in the additive heterogeneous diffusion model, and we also identify some sampling schemes that do not have this property. Our investigation may help to guide the design of studies when data on the entire population cannot be collected, thereby enabling the study of diffusion when social units are numerous or measurement of their characteristics is expensive.

2. MONTE CARLO PROCEDURES

We conducted all experiments by simulating diffusion histories governed by equation (1). To do this, we modified a Fortran program named EHG (Event History Generator),[2] which uses Monte Carlo techniques to generate event-history data. The society (i.e., the population) had N^* members, in which N^* varied from 100 to 600 across experiments but was constant over time for any given society. Each diffusion history started at time 0 and ended when every member of the society had had an event. In the model given by equation (1), every member of the population eventually has an event with probability 1, but for certain parameter values, some members of the population have the event much later than most others in the population.

Each society had a social network in which every individual member was proximate (or linked) to q randomly chosen other members. In most experiments, we chose q to be 3, following the design of Greve et al. (1995). However, to examine whether differences in network degree affect the results, we also performed some experiments in which q was 15. Within a given experiment, the probability that one member of the society was linked to another member was constant and statistically independent of all other linkages in the society. Cliques (i.e., subsets of the population with a high level of mutual linkages) could arise by chance but were not an intentional feature of our design.

A binary measure of the proximity of a pair of individuals in the society is simple; the pair has a linkage or it does not. Continuous measures of a pair's proximity tend to be more informative because they distinguish various degrees of closeness and social distance. Greve et al. (1995) found that social networks with continuous variation in proximity, which many network measures yield, are estimated with greater precision than the binary measure examined here. We expect, therefore, that our study of a binary measure of proximity accentuates problems resulting from incomplete data and that some problems found by our study might be reduced in magnitude (though possibly not eliminated) if a suitable continuous measure of proximity were used instead. Our use of a binary measure of proximity in this study is analogous to miners using a canary to detect potentially dangerous levels of noxious gases: It helps us locate problems.

The additive heterogeneous diffusion model in (1) allows variables describing heterogeneity in the vectors for propensity, susceptibility, and contagiousness of the members of the society. We included one variable in each of these three vectors; each variable was independently drawn from a standard Gaussian distribution with a mean of 0 and a variance of 1.

To generate each new diffusion history, pseudorandom techniques were used to draw a new realization of the social network in the society and new realizations of the covariates for all members of the society. Results reported below are based on generating and analyzing data on 1,000 societies subject to each of the conditions we studied. All diffusion histories use a single set of true parameter values, similar to the one used by Strang and Tuma (1993) and identical to one used by Greve et al. (1995). They chose these parameter values based on the coefficient estimates from reanalysis of the tetracycline diffusion data (Coleman, Katz, and Menzel 1966). Greve et al. examined how changes in true parameter values affected the coefficient estimates.

After generating the complete diffusion history for a society, we then imposed time aggregation, censoring, or some sampling scheme to create an incomplete data set. The expected sample size in the incomplete data set was 100, except where noted otherwise. When we studied the impacts of censoring, there were exactly 100 cases in the sample. We sorted the cases by their event times and analyzed either the first 100 cases that had the event (right censoring) or the last 100 (left censoring). In our studies of random sampling, we used Bernoulli sampling with p as the probability of choosing each member of the society, such that the expected sample size was a given number, usually 100. The number of cases in the randomly chosen sample had a binomial distribution. Thus, the actual sample size could be somewhat larger or smaller than its expected value.

In each experiment (i.e., for each condition studied), we generated and analyzed 1,000 diffusion histories using a version of the RATE computer program adapted to diffusion studies by Strang and Tuma (1993). To summarize the results, we recorded and report the means and standard deviations of all parameter estimates. To learn whether the estimated confidence intervals were consistent with the chosen level of significance, we also tallied the frequency with which the true value lay within the nominal two-sided 90 percent confidence interval of the estimated parameter. The nominal two-sided 90 percent confidence interval equals the estimated parameter plus/minus its estimated standard error multiplied by 1.645, the critical value based on a large-sample Gaussian approximation. We also tallied the frequency of rejections of the null hypothesis that a given parameter is zero using the 10 percent level of significance, but we do not report them because they were nearly always close to 100 percent. In conditions in which the frequency of correct rejection of the null hypothesis was low, we found that confidence intervals were also imprecise. We focus on these instead.

These experiments are designed to let us assess the model and estimation procedure under conditions similar to those encountered by many social scientists studying diffusion processes. Limiting the sample to 100 cases helps to establish the properties of the model and estimators when the number of cases analyzed is relatively small. Due to the lengthy computation time involved in simulating diffusion histories in large populations (see the appendix), we did not study populations with more than 600 cases. We emphasize that *simulation* of the diffusion histories of large populations is extremely time-consuming. Parameters of the model are readily estimated from much larger data sets than those analyzed in the studies reported below.

3. RESULTS

3.1. TIME AGGREGATION

Time aggregation occurs when the observation plan does not record event times exactly but with a certain degree of imprecision. To give an empirical example, archives may record the dates of events to the nearest day, month, or year and not the exact moment of occurrence. Such data tell only the time interval (e.g., which day, month, or year) in which an event occurred. Time aggregation can make truly sequential events appear to have occurred simultaneously, producing artifactual ties in event times.

Some degree of time aggregation usually occurs in the collection of event histories. For example, governmental decisions and organizational actions are often dated to the year, and reports of actor behavior by third parties (such as the press and legal agencies) often record events only when the source has noticed them. Actors' own reports of their recent events also vary in precision. Longitudinal (panel) surveys may record events intermittently because continuous observation is too costly (Coleman et al. 1966) or too burdensome on respondents.

Time aggregation in ordinary continuous-time parametric hazard-rate models has been studied by Petersen (1991) and Petersen and Koput (1992). They found that time aggregation leads to biased estimation when the observed times are treated as accurate and that the bias is large if the measurement interval is large relative to the average waiting time to an event. They learned that bias could be reduced by choosing the event time as the midpoint of the interval, or even earlier, with an earlier time being optimal when the hazard rate was high. They obtained unbiased estimates, however, only when using a likelihood that took into account the aggregation of event times.

These previous results are suggestive but do not transfer directly to diffusion analyses. In diffusion studies, time aggregation produces the additional problem of imprecise updating of the contagion component of the model. The diffusion model assumes that actors who have previously had events can affect the hazard rate of those who have not yet experienced the event. Thus, social influence runs strictly from earlier events to later events and not from future events to previous events. Consequently, a researcher must make some assumption about how actors whose event times appear to be tied did or did not influence each another. A priori it is not obvious which assumption might tend to minimize bias.

We explored the effect of time aggregation on parameter estimation by generating the diffusion history for a population of size 100 and then imposing different degrees of imprecision on the measured event times before analyzing the data. The diffusion history for a society was not censored, so an event time was observed for every member of the population. This procedure was repeated 1,000 times (i.e., for 1,000 societies) for each condition to examine the impact of random fluctuations in the network and in covariates' values on estimator quality and also on statistical inferences based on estimated standard errors of parameters. Each diffusion history had no tied event times in reality or in the finest time resolution that we studied, but roughly a third of all event times appeared to be tied in the crudest time resolution. Hence, the time aggregation that we studied ranged from negligible to extensive.

We report results of analyses that treated events as occurring at the *start* of the time interval in which they fell because we found that this choice led to better estimates than treating events as occurring at either the midpoint or endpoint of the time interval.[3] We then performed analyses making different assumptions about how cases with tied event times influenced each other. First, cases with tied events were assumed not to influence each other. Second, cases with tied events were allowed to influence each other, with the influence starting at the measured event time. This meant that cases with tied events could influence each other only for an instant. Third, cases with tied events were allowed to influence each other, with the influence starting at the measured event time lagged by one-half of the width of the time interval. The third assumption meant that cases with tied events could influence one another for a brief period of time. Since these three assumptions turned out to yield virtually identical results, we report only the estimates assuming that cases with tied events did not influence each other.

Table 1 reports the results from the estimation. Panel A shows that time aggregation leads to biased estimates of the parameters in the propensity vector. The standard deviations of these parameter estimates increase as the time resolution becomes cruder, indicating a loss of efficiency. Panel B points to problems of inference concerning parameters in the propensity vector. The nominal two-sided 90 percent confidence interval contains the true value in fewer than 90 percent of the 1,000 repetitions of the experiment in the case of the propensity vector, although roughly the correct percentage in the case of the effects of the susceptibility, contagion, infectiousness, and proximity variables. Panel C shows that under time aggregation, the percentage deviation of the estimate from the true value increases for parameters in the propensity vector but is not much different for the other parameters in the model.

Although the deterioration in the quality of estimation is apparent only for the propensity term in the model, it is large enough to imply that a high degree of time aggregation is undesirable. In other analyses, we found that setting event times to the midpoint or endpoint of the interval led to even worse estimates of the parameters in the propensity vector, with clear bias and overly narrow confidence intervals. These alternative schemes did not, however, adversely affect the quality of estimation of the parameters in the other vectors.

TABLE 1: Time-Aggregated Data

		Time Resolution				
Parameter	*True*	*0.0001*	*0.01*	*0.02*	*0.04*	*0.1*
Panel A: Mean ML estimate (standard deviation)						
Propensity intercept	–6.0	–6.0 (0.7)	–6.4 (1.1)	–6.4 (1.1)	–6.3 (1.3)	–5.8 (1.4)
Propensity covariate	5.0	5.0 (0.5)	5.4 (0.8)	5.5 (0.9)	5.4 (1.0)	5.1 (1.1)
Susceptibility	2.0	2.0 (0.1)	2.0 (0.1)	2.0 (0.1)	2.0 (0.1)	2.0 (0.1)
Contagion intercept	–8.0	–8.2 (0.7)	–8.1 (1.1)	–8.0 (0.7)	–8.1 (0.9)	–8.0 (1.1)
Infectiousness	2.0	2.1 (0.5)	2.1 (0.5)	2.0 (0.5)	2.0 (0.6)	2.0 (0.6)
Social proximity	4.0	4.1 (0.4)	4.0 (0.4)	4.0 (0.4)	4.0 (0.5)	3.9 (0.6)
Panel B: Percentage of true values within nominal 90 percent confidence interval						
Propensity intercept		90	79	75	70	57
Propensity covariate		90	72	66	63	53
Susceptibility		90	88	88	89	91
Contagion intercept		91	89	87	86	86
Infectiousness		90	90	88	86	86
Social proximity		91	92	91	91	91
Panel C: Percentage deviation from true value						
Propensity intercept		9	14	15	16	17
Propensity covariate		8	14	15	16	17
Susceptibility		5	6	6	6	6
Contagion intercept		8	7	7	7	8
Infectiousness		17	19	19	19	20
Social proximity		9	8	9	9	10

NOTE: ML = maximum likelihood.

Table 1 also indicates, however, that even substantial time aggregation produces no detectable bias in the estimates of contagion effects. The contagion intercept and the influence of susceptibility, infectiousness, and social proximity variables are all estimated accurately. This is a surprising and important finding because artifactual ties due to time aggregation are a common occurrence in empirical studies.

To understand the good estimation of the parameters in the susceptibility, contagion, infectiousness, and proximity terms, despite poor estimation of the parameters in the propensity term, it helps to view time aggregation as simultaneously coarsening the time scale and shifting the observed event times to the chosen point in the interval (the start of the interval in the results we report). The coarsening increases standard deviations of all parameter estimates, as one would expect to happen when information is lost. The shift in the event time causes bias in the estimates in the propensity vector, as it does in standard, nondiffusion event-history analyses (Petersen 1991; Petersen and Koput 1992). In contrast, estimates of the parameters in the other terms of the model rely mainly on the duration between an influential event and an influenced event. Since these event times are shifted by the same amount on average, there is no apparent bias in the estimated parameters in the social influence terms of the model.

Time aggregation impairs estimation of the propensity to have the event but does not appreciably bias the estimation of the effects of covariates in the susceptibility, infectiousness, and proximity terms. Estimates are more precise when there is less time aggregation and, other things being equal, when the starting point of the time interval is treated as the event time.

3.2. INCOMPLETE OBSERVATION ON THE RIGHT (RIGHT CENSORING)

As is well known, usually right censoring is relatively unproblematic when estimating hazard-rate models from event-history data. Even small samples with extensive right censoring can yield high-quality estimates provided the model is correctly specified (Tuma and Hannan 1978). One can hope for a similar result when estimating equation (1) from right-censored diffusion histories, but one cannot draw a clear conclusion a priori because of the model's unique features. In the diffusion model, cross-case interdependence may be difficult to capture if too few event times are observed due to right censoring.

To examine right censoring, we first generated a diffusion history for the entire society and then sorted members of the population by the times of their events. The times of the first 100 events were recorded exactly as generated. The later events of the other cases were treated as censored at the time of the 100th event. This censoring scheme, an example of what statisticians call Type II censoring (cf. Miller 1981), mimics the approach of a researcher who waits for the occurrence of the first 100 events and then censors the event times for the other cases. We then estimated the diffusion model in (1) from the resulting right-censored diffusion histories.

The estimates (see Table 2) are rather good. In every condition, the average estimates are close to the true values of the parameters. There is little sign of deterioration in estimator quality as the proportion of censored cases rises to 5/6, although there may be some deterioration for the contagion intercept and the coefficient of the infectiousness variable. The efficiency of the estimates does not seem to vary with the censoring proportion except for the coefficients of the infectiousness and social proximity variables, which are estimated less efficiently when the proportion of right-censored cases is high. The nominal 90 percent confidence interval contains the true value in close to 90 percent of the cases, implying that statistical inferences will be correct. *Parameters in the diffusion model can apparently be estimated well from right-censored histories.*

3.3. INCOMPLETE OBSERVATION ON THE LEFT (LEFT CENSORING AND LEFT TRUNCATION)

Left-censored and left-truncated event histories are troublesome (Tuma and Hannan 1984, chap. 5; Guo 1993; Wu 1996) even when there is no cross-case interdependence. We expect problems to be at least as bad in analyses of diffusion histories because failure to collect data on the early part of the history leads to two methodological problems. First, the cases that have had events before observation begins are truncated (not observed at all); second, the cases that have not yet had events are left censored (initially observed some time after they became at risk of the event). In particular, lack of data on early events may be more damaging when estimating diffusion models rather than ordinary event-history models because early unobserved events affect the timing of later observed events, causing a form of specification bias. Nevertheless, it is important to study estimator quality when analyzing diffusion histories that are left censored and/or left truncated, because these data problems are so common in certain social scientific

TABLE 2: Data With Incomplete Observation on the Right (data censored on the right)

	Propensity			*Contagion*		*Social*
Condition	*Intercept*	*Covariate*	*Susceptibility*	*Intercept*	*Infectiousness*	*Proximity*
Panel A: Mean ML estimate (standard deviation)						
True value	–6.0	5.0	2.0	–8.0	2.0	4.0
None censored	–6.0 (0.7)	5.0 (0.5)	2.0 (0.1)	–8.2 (0.7)	2.1 (0.5)	4.1 (0.4)
1/6 censored	–6.0 (0.7)	5.1 (0.5)	2.0 (0.2)	–8.2 (0.9)	2.1 (0.5)	4.1 (0.4)
1/3 censored	–6.1 (0.8)	5.1 (0.5)	2.0 (0.2)	–8.2 (0.9)	2.1 (0.5)	4.0 (0.4)
1/2 censored	–6.0 (0.8)	5.0 (0.5)	2.0 (0.2)	–8.3 (0.9)	2.1 (0.6)	4.1 (0.4)
2/3 censored	–6.0 (0.7)	5.0 (0.4)	2.0 (0.2)	–8.4 (1.9)	2.1 (1.0)	4.1 (0.6)
5/6 censored	–6.1 (0.7)	5.0 (0.4)	2.0 (0.2)	–8.8 (4.5)	2.3 (2.1)	4.0 (1.0)

NOTE: ML = maximum likelihood.

fields (e.g., organizational studies) in which diffusion processes are of great interest (see Kogut and Parkinson 1998).

To simulate data with these problems, we again generated the diffusion history of the entire society and sorted cases by the event times. Now, however, we recorded the true values of the covariates and the event times of the *last* 100 cases to have events. The likelihood function excluded data on the cases with events before the last 100 cases. We studied two conditions. In one, the last 100 observed cases were given start times equal to zero (the true time at which they became at risk). In the second, the censoring time in the society was treated as the start of the diffusion process for all cases. We show results only for the second condition in which the start time equals the censoring time because it yielded slightly better estimates than the first condition. The second condition results when researchers unintentionally collect left-truncated and left-censored data because they are unaware of earlier events.

Table 3 shows how estimator quality suffers under this scheme. Every parameter estimate is biased for every proportion of unobserved cases, and biases increase as this proportion increases. Standard deviations of the estimates also increase as the proportion of unobserved cases increases. In addition, we found that the nominal 90 percent confidence intervals for many estimated parameters were inaccurate. Only the effect of the susceptibility variable had a bias that could be called minor, and then only for the lowest proportions of unobserved cases. Finally, as the proportion of unobserved cases rises, the iterative search for estimates that maximize the logarithm of the likelihood function failed to converge in some instances.[4] *Parameters of the diffusion model, like parameters in event-history models in which there are no cross-case influences, are not estimated well from diffusion histories subject to left truncation and left censoring.*

TABLE 3: Data With Incomplete Observation on the Left (data truncated and censored on the left)

	Propensity			*Contagion*		*Social*
Condition	*Intercept*	*Covariate*	*Susceptibility*	*Intercept*	*Infectiousness*	*Proximity*
Panel A: Mean ML estimate (standard deviation)						
True value	–6.0	5.0	2.0	8.0	2.0	4.0
All observed	–6.0 (0.7)	5.0 (0.5)	2.0 (0.1)	–8.2 (0.7)	2.1 (0.5)	4.1 (0.4)
1/6 unobserved	–5.2 (1.6)	4.3 (1.8)	2.2 (0.2)	–7.0 (1.8)	1.4 (1.1)	2.9 (2.0)
1/3 unobserved	–4.7 (1.7)	3.6 (2.1)	2.4 (0.3)	–6.9 (6.7)	1.0 (3.3)	1.8 (3.3)
1/2 unobserved	–4.6 (2.6)	3.3 (3.1)	2.6 (0.4)	–6.2 (5.4)	0.7 (2.9)	0.6 (4.1)
2/3 unobserved[a]	–4.4 (3.2)	2.6 (2.8)	2.8 (0.5)	–6.3 (11.3)	0.9 (5.3)	–0.3 (5.7)
5/6 unobserved[b]	–4.5 (8.1)	2.3 (4.4)	3.0 (0.6)	–6.1 (25.6)	1.0 (8.0)	–1.6 (8.1)

NOTE: ML = maximum likelihood.
a. 1 of 1,000 repetitions failed to converge.
b. 6 of 1,000 repetitions failed to converge.

It may be possible to derive a correct likelihood function to use when estimating the diffusion model from left-truncated and left-censored event histories, as has been done for some other models (e.g., Guo 1993). However, the interdependence of early and late events in the diffusion process greatly complicates this task. Collection of data on all early events may be a more promising approach than modifying the likelihood function.

3.4. SIMPLE RANDOM SAMPLING

Studies may have data on only a subset of the population as the result of either a conscious data-collection scheme, an overly narrow definition of the population's boundaries, or lack of information on certain key variables (e.g., due to survey nonresponse). To understand consequences of the first, we studied the impact on estimator quality of some simple random sampling schemes. We begin with the fundamental situation in which a random sample was chosen from the population using a fixed sampling probability p for each member of the population. Thus, in this set of experiments, the sample size has a binomial distribution.[5]

3.4.1. Estimation Without Adjustments for Sampling

The simplest approach is to estimate the diffusion model from data on the sample without adjusting for the unobserved influence on the sampled cases of the nonsampled cases (i.e., those in the population but not in the sample). In this experiment, we generated the diffusion history for the entire society and then drew random samples of cases from the whole society. We estimated the parameters of the diffusion model using only the data on the sampled cases. One expects this approach to yield biased estimates because there are inaccurate measurements of some right-hand-side variables, namely, those pertaining to cross-case influences.

Table 4 summarizes the results of these experiments. Some estimated parameters are biased but not all. For the conditions that we studied, the estimated parameters in the propensity vector are unbiased. There is also no apparent bias in the estimated coefficient of the susceptibility variable. Presumably, these effects can be estimated well because they refer to measured characteristics of the actors at risk of the event rather than to the unmeasured characteristics of the nonsampled cases, some of which may have influenced those in the sample.

Conversely, the estimated effects of the infectiousness and social proximity variables are biased toward 0 and are difficult to detect when the sampling probability p is small. Some such bias is apparent even when p is as large as 1/2, and the bias increases as p decreases. For example, when p is 1/6, the effects of the infectiousness and proximity variables are correctly detected as differing from 0 in only half of the repetitions of the experiment. The downward biases in these estimated effects are offset by an upward bias in the estimate of the contagion intercept. The true value of the contagion intercept is –8, and it is slightly underestimated using data on the whole population. But when p is 1/6, the average estimate is –4.8, which means that the estimate overstates the true value by a factor of 24.5 = exp(3.2) = exp(–4.8)/exp(–8).

TABLE 4: Random Sampling With No Adjustments

	Propensity					
Condition	*Intercept*	*Covariate*	*Susceptibility*	*Contagion Intercept*	*Infectiousness*	*Social Proximity*
Panel A: Mean ML estimate (standard deviation)						
True value	–6.0	5.0	2.0	–8.0	2.0	4.0
100 from 100	–6.0 (0.7)	5.0 (0.5)	2.0 (0.1)	–8.2 (0.7)	2.1 (0.5)	4.1 (0.4)
100 from 150	–6.0 (0.9)	5.0 (0.6)	2.0 (0.2)	–7.6 (2.7)	1.9 (1.3)	3.5 (0.9)
100 from 200	–6.0 (0.9)	5.0 (0.6)	2.0 (0.2)	–7.2 (5.8)	1.8 (2.9)	3.1 (1.5)
100 from 300	–6.1 (1.1)	5.1 (0.8)	2.0 (0.2)	–5.8 (3.2)	1.1 (1.5)	1.2 (5.1)
100 from 600	–6.2 (1.4)	5.2 (0.9)	2.0 (0.2)	–4.8 (1.5)	0.4 (1.3)	–1.5 (9.2)

NOTE: ML = maximum likelihood.

Moreover, there are inferential problems when estimating the diffusion model from a random sample without making any adjustments. The percentage of true values within the nominal 90 percent confidence intervals is consistently less than 90 percent, and it decreases as p is reduced. The decreases are especially marked for the contagion intercept and the effect of the infectiousness variable.

In sum, estimation of the diffusion model from a random sample of cases without any adjustments for sampling yields unbiased estimates of propensity and susceptibility effects but biased estimates of effects of infectiousness and social proximity. The extent of the bias increases as the sampling probability decreases.

3.4.2. Estimation With Supplementary Data on Nonsampled Cases

The bias found above results from the unmeasured influence of events occurring to nonsampled cases. Intrapopulation influences are a key aspect of diffusion processes, as well as some other social processes, such as competition for common resources (e.g., Hannan and Carroll 1992). Thus, a method for satisfactorily adjusting for the influence of nonsampled cases on the sampled cases not only is necessary for research on diffusion but may also suggest extensions to research on other kinds of cross-case influences.

One possible remedy is to adjust the estimation procedure to try to account for sampling. We explored several possibilities but focus here on a random sampling with supplementary data approach, the only strategy we considered that proved fairly successful. In this approach, we obtained supplementary information about the nonsampled cases, in particular, their event times, infectiousness, and proximity to the sampled members (but not their own propensity, susceptibility, or proximity to other nonsampled cases). The rationale for this approach is that information on nonsampled cases is needed only to the extent that it contributes to the likelihood function formed for the sampled cases.

In some contexts, collection of supplementary data on nonsampled cases is as burdensome as collecting complete data on all covariates for the entire population. This approach becomes attractive, however, where social proximity, infectiousness, and event times can be measured fairly easily, while attributes relating to propensity and susceptibility are hard to obtain. For example, in studies of business organizations, common measures of social proximity (e.g., ownership ties, directorship ties, and geographical location) are readily available, but organizational characteristics affecting the intrinsic propensity to adopt some practice (e.g., a new organizational strategy) can be very costly to measure. It is then sensible to limit the expensive data collection (e.g., measurement of propensity variables) to a sample but collect data on proximity and event times for the entire population.[6]

The results of this experiment are given for eight conditions, reported in the rows of Table 5. The first three rows give the results for complete data on the whole society for three different population sizes: 100, 200, and 300. These results provide a basis of comparison for other conditions in which the model is estimated from random samples of similar sizes. The next three rows give results for different population sizes when the sampling probability was one-half so that the expected sample size was 100, 200, and 300, respectively. The estimated parameters are less precise for each of the three sampled conditions than for an entire population of a similar size. Still, the estimated parameters based on the random samples are fairly close to the true values and have only slightly larger standard deviations than in the comparable similar-sized population.[7] (Compare results in rows 4 and 1, rows 5 and 2, and rows 6 and 3.) As the population and sample sizes increase, deviations from the true value decline (panel C), as do estimated standard errors (not shown). The percentage of estimates lying within the nominal 90 percent confidence interval is just below 90 percent for most parameters in both the full population and the sampled conditions (panel B). The sampled conditions yield somewhat lower percentages of estimates within the nominal 90 percent

TABLE 5: Random Sampling With Supplementary Data

	Propensity					
Condition	*Intercept*	*Covariate*	*Susceptibility*	*Contagion Intercept*	*Infectiousness*	*Social Proximity*
Panel A: Mean ML estimate (standard deviation)						
True value	–6.0	5.0	2.0	–8.0	2.0	4.0
100 from 100	–6.0 (0.7)	5.0 (0.5)	2.0 (0.1)	–8.2 (0.7)	2.1 (0.5)	4.1 (0.4)
200 from 200	–6.0 (0.5)	5.0 (0.4)	2.0 (0.1)	–8.1 (0.6)	2.0 (0.4)	4.0 (0.3)
300 from 300	–6.0 (0.5)	5.0 (0.3)	2.0 (0.1)	–8.1 (0.5)	2.0 (0.3)	4.0 (0.3)
100 from 200	–6.1 (0.9)	5.1 (0.6)	2.0 (0.1)	–8.2 (1.1)	2.1 (0.6)	4.1 (0.5)
200 from 400	–6.1 (0.7)	5.1 (0.4)	2.0 (0.1)	–8.2 (0.8)	2.1 (0.5)	4.1 (0.4)
300 from 600	–6.1 (0.6)	5.1 (0.4)	2.0 (0.1)	–8.1 (0.7)	2.0 (0.4)	4.0 (0.4)
100 from 300	–6.2 (1.1)	5.2 (0.7)	2.0 (0.1)	–8.3 (1.2)	2.1 (0.7)	4.1 (0.7)
100 from 600	–6.3 (1.5)	5.2 (1.0)	2.0 (0.1)	–8.5 (3.1)	2.2 (1.3)	4.0 (1.2)
Panel B: Percentage of true values within nominal 90 percent confidence interval						
100 from 100	90	90	90	91	90	91
200 from 200	89	89	90	88	87	90
300 from 300	89	90	90	88	89	90
100 from 200	88	87	89	86	87	89
200 from 400	90	90	90	86	87	88
300 from 600	91	90	89	86	86	89
100 from 300	88	87	91	87	87	89
100 from 600	88	87	90	83	84	88
Panel C: Percentage deviation from true value						
100 from 100	9	8	5	6	17	8
200 from 200	7	6	4	6	15	6
300 from 300	6	5	3	5	13	6
100 from 200	12	9	6	9	22	10
200 from 400	9	7	4	7	17	8
300 from 600	7	6	3	7	16	7
100 from 300	13	10	5	10	24	12
100 from 600	16	13	6	15	32	16

NOTE: ML = maximum likelihood.

confidence interval for the contagion intercept and the infectiousness variable, however.

The last two rows give results for random samples with an expected size of 100 that are drawn from populations of 300 and 600, respectively. One should compare the results in rows 7 and 8 with those in row 4 and all three of these rows with the results for the similar-sized population in row 1. As the sampling probability *p* decreases, the estimates become less efficient even though the expected sample size is the same. The sample of 100 from a society of 600 has rather imprecise parameter estimates as shown by the high percent deviation from the true value (panel C). This result suggests that low sampling probabilities should be avoided.

Random sampling with inclusion of data on the event times, infectiousness variables, and social proximity of nonsampled cases that have had events yields unbiased estimates. The efficiency of these estimates is lower than for data on an entire population of a similar size, and it decreases as the sampling probability decreases.

3.4.3. Estimation From Samples of High-Degree Networks

When estimating a heterogeneous diffusion model, a particular concern is whether characteristics of the social network in the population affect the bias and efficiency of parameter estimates. Results using data on the entire population (Greve et al. 1995) suggest that continuously valued network measures, such as structural equivalence (Burt 1980), yield better estimates than the dichotomous measures of proximity used here and that high-degree networks (i.e., ones in which actors have many linkages) yield less efficient estimates than low-degree networks (i.e., ones in which actors have few linkages). These findings point to the value of determining if the above results hold when individuals have many linkages. In the next set of experiments, we generated networks in which each member was socially proximate (i.e., linked) to 15 randomly chosen others. We studied the same eight conditions reported in Table 5. The results are given in Table 6.

Conditions based on data for the entire population of a given size yield estimates that are no more biased in the 15-tie network (Table 6) than in the 3-tie network (Table 5). But the estimates in the 15-tie network do tend to have larger standard deviations. The differences between the results for the 15-tie and 3-tie networks decrease as the population size increases.

TABLE 6: Random Sampling From a High-Degree Social Network

	Propensity			*Contagion*		*Social*
Condition	*Intercept*	*Covariate*	*Susceptibility*	*Intercept*	*Infectiousness*	*Proximity*
Panel A: Mean ML estimate (standard deviation)						
True value	–6.0	5.0	2.0	–8.0	2.0	4.0
100 from 100	–6.1 (1.1)	5.1 (0.7)	2.0 (0.1)	–8.6 (2.3)	2.1 (0.4)	4.5 (2.5)
200 from 200	–6.0 (0.7)	5.0 (0.4)	2.0 (0.1)	–8.0 (0.7)	2.0 (0.3)	4.0 (0.7)
300 from 300	–6.0 (0.6)	5.0 (0.4)	2.0 (0.1)	–8.0 (0.3)	2.0 (0.2)	4.0 (0.3)
100 from 200	–6.1 (1.1)	5.1 (0.8)	2.0 (0.1)	–8.3 (1.4)	2.1 (0.4)	4.2 (1.6)
200 from 400	–6.1 (0.8)	5.1 (0.5)	2.0 (0.1)	–8.0 (0.5)	2.0 (0.3)	4.0 (0.5)
300 from 600	–6.1 (0.6)	5.1 (0.4)	2.0 (0.1)	–8.0 (0.4)	2.0 (0.3)	4.0 (0.3)
100 from 300	–6.4 (2.6)	5.3 (1.4)	2.0 (0.1)	–8.4 (1.4)	2.1 (0.5)	4.2 (1.5)
100 from 600	–6.4 (2.0)	5.3 (1.3)	2.0 (0.1)	–8.4 (1.6)	2.1 (0.7)	4.2 (1.1)

NOTE: ML = maximum likelihood.

In conditions estimated from data on the sampled cases plus supplementary data on the nonsampled cases' event times, proximity, and infectiousness, the standard deviations of the estimated parameters in the propensity vector are larger for the 15-tie network (Table 6) than for the 3-tie network (Table 5). However, the standard deviations of most of the estimated parameters pertaining to intrapopulation influences are smaller for the 15-tie network than the 3-tie network, especially when the expected sample size is greater than 100, suggesting that coefficient estimates of social influence based on sampled data are more efficient in high-degree networks in which actors have many links.

Small populations with high-degree networks cause problems in estimating the contagion intercept and proximity effect. However, random sampling with inclusion of data on the event times, infectiousness variables, and social proximity for the nonsampled cases satisfactorily corrects biases due to sampling. Moreover, for a given expected sample size, estimates of the contagion intercept and the social proximity effect are more efficient when the population is larger.

3.5. ESTIMATION FROM OTHER SAMPLING PLANS

We have found that the quality of estimates using supplementary data on nonsampled cases is good when the sampling probability is the same for all members of the population. This sampling design is, however, only one of many, and it is among the simplest. Other common random sampling designs include clustered sampling, stratified sampling, and sequential sampling (e.g., see Kish 1965), as well as still others that combine selected features of these designs. These various intentional sampling designs are usually employed to achieve greater cost-effectiveness.

To complicate matters further, the implementation of a particular intentional sampling design often yields an *observed* sample that lacks the intended properties because data on certain members of the target sample are missing either entirely or for certain key variables. The likelihood that there are complete data on a member of the population may be related to the explanatory covariates, the outcome being studied, or both.

Whatever the intentional sampling design, the observed sample in sociological studies of diffusion is more likely to omit less prominent actors, such as less-developed nations, small firms, or low-status members of a community. A researcher's difficulty in obtaining data on less prominent actors may be mirrored by inattention to these actors by the other members of the population. In this case, the infectiousness of the observed sample members tends to be greater on average than the infectiousness of the entire population. In contrast, in epidemiological research, less prominent actors may be more infectious than the population as a whole, leading infectiousness to be lower in a sample than in the population. Likewise, the sample distribution of other independent variables may also differ from the population-level distribution of these same variables. In practice, researchers rarely know how the distribution of independent variables in a sample differs from the distribution in the population.

Full treatment of complex sampling designs and patterns of missing data lies outside the scope of the present study. However, we considered three sampling schemes that are somewhat more complicated than the simple random sampling scheme considered in section 3.4. In all these experiments, we continue to include supplementary data on nonsampled cases.

We first studied a simple stratified sampling design in which a variable used to stratify the sample was uncorrelated with the independent variables affecting the diffusion process. The estimates were comparable in quality to those for samples chosen randomly with a fixed sampling probability. Since empirical research rarely employs a stratified sampling design in which the stratification variables are completely unrelated to all explanatory variables, we do not report the numerical results here, but they are available from the authors on request.

We next performed the following Monte Carlo experiment. First we generated a pseudorandom standard Gaussian variable Z that had correlation $\rho = 0.7$ with one of the covariates affecting the diffusion process. Then, we included a member of the population in our sample if the realization of Z exceeded a certain value $z^{\dagger}$ chosen to give a certain expected sample size. In our experiments, we chose the population size to be 300 and the expected sample size to be 100, implying $z^{\dagger} = 0.43$.

We studied the three conditions in which Z is correlated with the propensity, susceptibility, or infectiousness covariate, respectively. In each condition, Z was correlated with only one covariate and was uncorrelated with the other two covariates in the model. For each condition, we generated diffusion histories for 1,000 societies and estimated the model using the same likelihood function as before, thereby ignoring the sampling design.

Table 7 reports the results. There are no particular problems with either point estimates or inferences based on estimated standard errors in the conditions in which the sampling process is correlated with contagion effects. The estimated intercepts in the correlated sampling designs are slightly lower than the true values, but the same pattern also occurs in the uncorrelated sampling design with a fixed probability (cf. Table 5). The percentage of true values within the nominal 90 percent confidence interval is usually slightly below the desired value of 90 percent, implying that the nominal confidence intervals are slightly too narrow.

TABLE 7: Sampling Probability Correlated With a Covariate

	Propensity					
Condition	*Intercept*	*Covariate*	*Susceptibility*	*Contagion Intercept*	*Infectiousness*	*Social Proximity*
Panel A: Mean ML estimate (standard deviation)						
True value	–6.0	5.0	2.0	–8.0	2.0	4.0
No correlation	–6.2 (1.1)	5.2 (0.7)	2.0 (0.1)	–8.3 (1.2)	2.1 (0.7)	4.1 (0.7)
Propensity, $\rho = 0.7$	–6.0 (0.6)	5.0 (0.4)	2.0 (0.2)	–8.2 (1.2)	2.1 (0.7)	4.1 (0.6)
Propensity, $\rho = -0.7$	–8.2 (9.6)	7.0 (10.9)	2.0 (0.1)	–8.3 (1.3)	2.1 (0.7)	4.1 (0.5)
Susceptibility, $\rho = 0.7$	–6.2 (1.3)	5.1 (0.8)	2.0 (0.2)	–8.2 (1.0)	2.1 (0.6)	4.1 (0.5)
Susceptibility, $\rho = -0.7$	–6.1 (0.8)	5.1 (0.5)	2.0 (0.2)	–8.4 (1.3)	2.1 (0.7)	4.2 (0.6)
Infectiousness, $\rho = 0.7$	–6.1 (1.0)	5.1 (0.6)	2.0 (0.1)	–8.2 (1.0)	2.1 (0.6)	4.1 (0.5)
Infectiousness, $\rho = -0.7$	–6.1 (1.0)	5.1 (0.7)	2.0 (0.1)	–8.3 (1.2)	2.1 (0.6)	4.1 (0.6)

NOTE: Population size = 300; expected sample size = 100. ML = maximum likelihood.

One condition does appear troublesome, however. When Z is negatively correlated with the propensity covariate, the efficiency of both parameter estimates in the propensity term is very low. In this condition, the nominal 90 percent confidence intervals for both parameters are too narrow: The true values for the propensity intercept and the effect of the propensity covariate lie within their nominal 90 percent confidence intervals in fewer than 80 percent of the repetitions of the experiment. Moreover, the estimates of the two parameters in the propensity vector may be biased, although it is hard to be certain because the standard deviations of these two parameter estimates are so large.

The survivor function estimated from the sample data provides a helpful clue to these estimation problems. The exclusion of high-propensity cases causes the estimated survivor function to have a characteristic pattern of upward bias that starts very early. Since high-propensity cases have large hazard rates relative to others in the population, especially early in the process, omitting these actors from the sample causes a dearth of early observed events, much as occurs in left-truncated and left-censored data, which we found to be estimated with bias (section 3.3). By contrast, sampling on a variable correlated with the susceptibility covariate mainly biases the estimated survivor function at later times, and sampling on a variable correlated with the infectiousness variable does not appear to bias the survivor function much at all.

To study the problems that may arise when sampling depends even more strongly on the explanatory variables, we next examine samples chosen by omitting cases when one of the explanatory variables has values outside a specified interval. This more extreme sampling scheme is designed to let us explore a form of "worst-case scenario" rather than to examine a common empirical practice.

For this set of experiments, we again chose an expected sample of size 100 from a population of 300 by including all members whose values on one of the explanatory variables were either high ($z > 0.43$), medium ($-0.43 \leq z \leq 0.43$), or low ($z < -0.43$). These values are the ones that divide the cumulative probability distribution function for a standard Gaussian random variable into thirds. (Recall that in all of our experiments, the covariates were generated to have a standard Gaussian distribution.)

The results are given in Table 8. We again find that sampling processes related to contagion effects yield good estimates, although standard errors are somewhat higher than in the situation of equal-probability sampling (Table 5). But truncation of the sample based on intrinsic propensities yields poor estimates. Retention of only moderate- or low-propensity actors leads to biased and inefficient estimation of propensity effects. And retention of only high-propensity actors leads to inefficient estimation of contagion effects, presumably because these effects do not contribute substantially to outcomes in this subset of the population.

We thus find that diffusion estimates are surprisingly robust when sampling probabilities are related to variables affecting the contagion component of the model, but not when sampling is related to propensity effects. In particular, *accurate data on the early period of the diffusion process appear particularly important for good estimation of the diffusion model.*

4. IMPLICATIONS FOR DATA-COLLECTION STRATEGIES

The implications of our results for data-collection strategies for diffusion studies are clear. Leaving aside issues surrounding sampling, the data requirements for obtaining good estimates of the additive

TABLE 8: Truncated Sampling

	Propensity					
Condition	*Intercept*	*Covariate*	*Susceptibility*	*Contagion Intercept*	*Infectiousness*	*Social Proximity*
Panel A: Mean ML estimate (standard deviation)						
True value	–6.0	5.0	2.0	–8.0	2.0	4.0
Propensity high	–6.1 (0.6)	5.0 (0.4)	2.1 (0.3)	–8.6 (2.4)	2.2 (1.2)	4.2 (1.2)
Propensity medium	–10.3 (13.4)	13.7 (33.9)	2.1 (0.2)	–8.3 (1.0)	2.1 (0.6)	4.1 (0.5)
Propensity low	–0.3 (18.1)	14.9 (33.8)	2.1 (0.1)	–8.3 (1.0)	2.1 (0.6)	4.1 (0.5)
Susceptibility high	–6.3 (2.0)	5.2 (1.2)	2.0 (0.2)	–8.3 (1.2)	2.1 (0.7)	4.1 (0.6)
Susceptibility medium	–6.3 (1.2)	5.2 (0.8)	2.0 (0.5)	–8.4 (1.5)	2.1 (0.8)	4.2 (0.6)
Susceptibility low	–6.0 (0.6)	5.1 (0.5)	2.0 (0.3)	–8.9 (3.2)	2.3 (1.3)	4.4 (1.4)
Infectiousness high	–6.2 (1.2)	5.1 (0.7)	2.0 (0.1)	–8.2 (1.1)	2.0 (0.7)	4.1 (0.7)
Infectiousness medium	–6.1 (0.9)	5.1 (0.6)	2.0 (0.2)	–8.4 (1.5)	2.1 (0.7)	4.2 (0.7)
Infectiousness low	–6.1 (1.0)	5.1 (0.7)	2.0 (0.1)	–8.3 (1.4)	2.1 (0.7)	4.1 (0.7)

NOTE: Population size = 300; expected sample size = 100. ML = maximum likelihood.

heterogeneous diffusion model in equation (1) are largely those for other event-history models that lack cross-case influences. Time aggregation should be minimized to restrict bias in estimated effects on intrinsic propensities to have the event, but we discovered no special problems in estimating contagion effects due to the artifactual ties that result from time aggregation. And, while right censoring is relatively unproblematic, left truncation and left censoring do cause bias and should be avoided.

Sampling poses special problems for diffusion research, however. Good estimates of propensity and susceptibility effects may be obtained from samples, which is reassuring for researchers who are only interested in these effects. But estimates of the effects of infectiousness and social proximity are downwardly biased in random samples consisting of as many as half of the complete population. Our study indicates that supplementary data should be obtained if one wishes to obtain good estimates of the full heterogeneous diffusion model from a random sample. Namely, in addition to all relevant data on a random sample, one should also collect data on event times and the values of infectiousness and proximity variables for the nonsampled cases that have had events.[8] Although the supplementary data are used only as right-hand-side variables, including them minimizes biases in estimated effects of variables in the susceptibility and contagion components of the model.

One further issue should be noted, however. Throughout this discussion, we have assumed that the locations of the boundaries of the relevant population are known, but this assumption is often problematic (cf. Thornton and Tuma 1995). Although this assumption may be approximately true for many research projects, population boundaries are sometimes uncertain, and the degree of uncertainty may vary widely. At one extreme, uncertainty may extend to only a few cases; at the other, the researcher may not know whether geographical borders for diffusion of some phenomenon lie at the city, state, national, or global level. The conduct and costs of a study vary radically depending on the location of the population's boundaries. This fact points to the value of some formal procedure for determining the location of the actual boundaries of a population in which cross-case influences have effects.

Our procedure of random sampling with supplementary data (section 3.4.2) can be used as the basis of such a formal procedure. Collection of data on the favored, narrower specification of the population and on event times, infectiousness, and social proximity for other cases outside this specified population can be used to test whether inclusion of the data on the latter cases affects results based on analysis of data on the former cases. This is done by estimating the model once using the favored, narrower specification of the population and then again using the broader definition and the procedure that adjusts for sampling (cf. section 3.4.2). The infectiousness vector of the second model should include an indicator variable set to 1 for observations outside the narrowly defined sample. The test is based on the estimated coefficient of this indicator variable. This estimate should be a large negative number (implying low influence) if the narrow specification of the population is correct. This procedure is heuristic because the narrower specification of the population is not a random sample of the larger one. Nevertheless, it is likely to be more informative than the common approach of defining population boundaries without testing whether the chosen definition encompasses the actual population boundaries.

In large social networks, an additional boundary problem is whether diffusion involves influence only among actors that are directly tied to each other or whether there is also influence among actors with no direct ties. This boundary problem also influences data collection costs since influence among actors with no direct ties calls for data on all events in the population. In contrast, influence solely from connected members of the population requires only data on a sample and the actors directly linked to sample members. Previously, we showed that a similar procedure of estimating models with both an "everybody-influential" specification and a "only-ties influential" specification helps an analyst choose between these two specifications (Greve et al. 1995:402-03).

5. CONCLUSION

Our investigation shows that the additive heterogeneous diffusion model in equation (1) can be estimated well from data with event times that are measured precisely (but not those measured imprecisely), data with right censoring (but not left truncation and left censoring), and data drawn as an equal-probability sample from a larger population—provided one also includes data on event times, social proximity, and infectiousness variables for nonsampled cases that have had events. Since these requirements for data collection are often feasible, they mean that the model can be applied in a variety of empirical settings. The requirement that some data on nonsampled cases be collected is onerous but hard to avoid, given that *intrapopulation influences are central to diffusion studies*. Indeed, it would be surprising if one could estimate such models well without observing the acts of most cases that influence sample members.

As a practical matter, the problem of locating influential actors can be solved in two different ways. First, one can measure event times in the population and then collect data on linkages between sampled cases and nonsampled cases that have already had the event. Second, one can collect data on the networks of cases in some target sample and then collect data on event times from those nominated as socially proximate to those in the original target sample, as in snowball sampling procedures. For the study of very large populations, these data requirements tend to restrict diffusion studies to situations in which

few actors have had events (so that it is feasible to obtain data on all cases that have had events) or situations in which diffusion is strongly constrained by network ties (so that the total number of actors influencing members of the original sample is not too large). While neither condition may be met in some contexts, many large-scale diffusion processes can be investigated using one or the other of these two data collection strategies.

The ability to estimate the parameters of diffusion processes from sampled populations is likely to be extremely important for progress in the study of social influence. The requirement that one collect complete population data limits not only the sizes of populations that can feasibly be studied but also the kinds of populations that can be studied. Small villages and neighborhoods are thus popular settings for diffusion studies of individuals (Rogers 1995), while research on diffusion in larger social systems has seemed not only daunting but also infeasible. Focused, small-system studies have considerable theoretical and empirical value, but it is important to examine diffusion processes occurring in large populations as well. The results of the simulation experiments reported in this article suggest that the scale and scope of diffusion studies can be larger than most previous researchers have presumed.

APPENDIX
Monte Carlo Simulation Procedure

The programs for simulating heterogeneous event histories in general and diffusion histories in particular (EHG) and for analyzing them (RATE), along with the user manuals, can be found on the Web at www.stanford.edu/~tuma. A brief description of the algorithm for simulating the event times is given below.

All covariates were generated under the assumption that they have a standard Gaussian distribution with mean 0 and variance 1. Each realization of a covariate was drawn independently, except as noted otherwise in the text. We used the GGNML routine of the IMSL library for this purpose.

To draw each socially proximate observation, we first used the GGUBS routine of the IMSL library to generate values from a uniform distribution. Then we multiplied it by N^* and rounded up to the nearest integer, yielding the observation number of the proximate case. The draws were independent except we made a new draw if the same actor was drawn twice.

In our study of random sampling, we used the GGBN routine of the IMSL library to draw Bernoulli (dummy) variables for the sampling indicators.

We used the following algorithm to simulate a diffusion history for a society of size N^* governed by the additive heterogeneous diffusion model in equation (1). The process starts at time $t = 0$ and begins with the computation of the propensity term of the model, $\exp(\alpha' x_n)$, and the susceptibility term, $\exp(\beta' v_n)$, for every member n of the society. These values are saved and used as needed in subsequent steps until the complete diffusion history for the society has been generated.

Once the first case has experienced the event, the diffusion history for a society is generated using the following steps in an iterative fashion until every member of the society has had the event. The procedure is continued until all members of the society have had the event because the heterogeneous diffusion model implies that eventually every member of the society has the event. (The time of the last event in the society can, however, be very large.)

After i cases have had the event, with the most recent event being at some time t_i, do the following for each case n of the $N^* - i$ cases remaining in $\mathcal{N}(t_i)$:

1. Compute the value of $\Sigma_{s\in S(t_i)} \exp(\gamma' w_s + \delta' z_{ns})$.
2. Draw a realization from a uniform [0, 1] distribution; label it u_n. We used the GGUBS pseudorandom uniform number routine of the IMSL library for this purpose.
3. Find the value of the waiting time to the next event, w_n, such that $F(w_n) = u_n$, where $F(w)$ is the cumulative probability distribution function of the heterogeneous diffusion model; see the equations below. (The FORTRAN code for the inversion is available on the Web at the address given above.)

After the waiting times have been found for all cases in $\mathcal{N}(t_i)$, identify the case with the smallest waiting time. This case becomes the $(i + 1)$th case to experience the event. Its waiting time (found above) is relabeled w_{i+1}, and its event time is calculated as $t_{i+1} = t_i + w_{i+1}$. Move this case $i + 1$ from $\mathcal{N}(t_i)$ to $S(t_{i+1})$ and repeat the above steps for the remaining $N^* - (i + 1)$ cases that are still at risk at the new event time t_{i+1}.

This procedure is used because each case that has already had the event influences the remaining cases that have not yet had the event. Consequently, the social influence term of the model needs to be recomputed after each new event. One cannot compute event times for all members of the society in a single pass of N^* calculations, which is possible in models without social influence. Instead, after each new event, one needs to compute the waiting time to the next event for all cases still at risk and continue iteratively in this fashion until all cases have had the event. Therefore one needs to calculate $N^*(N^* - 1)(N^* - 2)\cdots 1 = (N^*)!$ waiting times to get the diffusion history for a society with N^* members. (The need to calculate $(N^*)!$ waiting times is the reason that it is very time-consuming to simulate the diffusion history of a very large society.)

In step 3, $F(w)$ has the usual definition

$$F_{(w)} = 1 - \exp\left(-\int_0^w r(\tau)d\tau\right),$$

and the waiting time for case n, w_n, is found as

$$w_n = -\frac{\log u_n}{\exp(\alpha' x_n) + \exp(\beta' v_n)\Sigma_{s\in S(t)} \exp(\gamma' w_s + \delta' z_{ns})}$$

NOTES

[1] The regularity conditions are discussed in Theil (1971) and can be informally stated as follows: (1) The first three derivatives of the log likelihood function with respect to the parameters are finite for all parameter values and almost all data values. (2) Expectations of first and second derivatives of the log likelihood function can be obtained. (3) The third derivative of the log likelihood function is less than a function that has a finite expectation.

[2] Event History Generator (EHG) was originally written by James C. Crutchfield under Nancy Tuma's guidance and later extended by Eric Bloch. Eric Bloch and Henrich Greve developed the module that generates diffusion histories.

[3] An interval estimator for the diffusion model has not yet been implemented. Its development requires decisions about what to assume concerning influences among cases with tied event times.

[4] We omitted runs that did not converge in computing the other statistics in Table 3.

[5] The normal approximation to the binomial distribution can be used to estimate the chances of drawing various sample sizes. For example, if $p = .5$ and the population size is 200, the expected sample size is $200(0.5) = 100$, and the standard deviation of the sample size is $\sqrt{200(0.5)(0.5)} \approx 7$. Then, roughly 95 percent of the samples are within the range 86 to 114.

[6] For applications of this approach in organizational studies, see Greve (1995, 1996) and Davis and Greve (1997).

[7] In Bernoulli sampling, many sample sizes are smaller than the average. The smaller than average samples seem to cause most of the imprecision in the estimates.

[8] Some network measures require complete data on the social network in the whole population. This requirement is familiar to social network researchers (see reviews such as Burt 1980; Bonacich 1987; Marsden 1990; Borgatti and Everett 1992). Data on a sample suffice in measuring certain aggregate network properties, such as the density of ties or the proportion of actors with a certain characteristic (Granovetter 1976; Frank 1981; Frank and Snijders 1994).

REFERENCES

Bonacich, Phillip. 1987. "Power and Centrality: A Family of Measures." *American Journal of Sociology* 92:1170-82.

Borgatti, Stephen P. and Martin G. Everett. 1992. "Notions of Position in Social Network Analysis." Pp. 1-35 in *Sociological Methodology*, edited by Peter V. Marsden. Cambridge, MA: Basil Blackwell.

Burt, Ronald S. 1980. "Models of Network Structure." *Annual Review of Sociology* 6:79-141.

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

Davis, Gerald F. and Henrich R. Greve. 1997. "Corporate Elite Networks and Governance Changes in the 1980s." *American Journal of Sociology* 103:1-37.

Frank, Ove. 1981. "A Survey of Statistical Methods for Graph Analysis." Pp. 110-55 in *Sociological Methodology*, edited by Samuel Leinhardt. San Francisco: Jossey-Bass.

Frank, Ove and Tom Snijders. 1994. "Estimating the Size of Hidden Populations Using Snowball Sampling." *Journal of Official Statistics* 10:53-67.

Granovetter, Mark. 1976. "Network Sampling: Some First Steps." *American Journal of Sociology* 81:1287-303.

Greve, Henrich R. 1995. "Jumping Ship: The Diffusion of Strategy Abandonment." *Administrative Science Quarterly* 40:444-73.

———. 1996. "Patterns of Competition: The Diffusion of a Market Position in Radio Broadcasting." *Administrative Science Quarterly* 41:29-60.

———. 1998. "Managerial Cognition and the Mimetic Adoption of Market Positions: What You See is What You Do." *Strategic Management Journal* 19:967-88.

Greve, Henrich R., David Strang, and Nancy Brandon Tuma. 1995. "Specification and Estimation of Heterogeneous Diffusion Models." Pp. 377-420 in *Sociological Methodology*, edited by Peter V. Marsden. Cambridge, MA: Basil Blackwell.

Guo, Guang. 1993. "Event-History Analysis for Left-Truncated Data." Pp. 217-43 in *Sociological Methodology*, edited by Peter V. Marsden. Cambridge, MA: Basil Blackwell.

Hannan, Michael T. and Glenn R. Carroll. 1992. *Dynamics of Organizational Populations*. Oxford, UK: Oxford University Press.

Kish, Leslie. 1965. *Survey Sampling*. New York: John Wiley.

Kogut, Bruce and David Parkinson. 1998. "Adoption of the Multidivisional Structure: Analyzing History from the Start." *Industrial and Corporate Change* 7:249-73.

Marsden, Peter V. 1990. "Network Data and Measurement." *Annual Review of Sociology* 16:435-63.

Marsden, Peter V. and Joel Podolny. 1990. "Dynamic Analysis of Network Diffusion Processes." Pp. 197-214 in *Social Networks Through Time*, edited by H. Flap and J. Weesie. Utrecht, the Netherlands: ISOR, University of Utrecht.

Miller, Rubert G., Jr. 1981. *Survival Analysis*. New York: John Wiley.

Myers, Daniel J. 1997. "Racial Rioting in the 1960s: An Event History Analysis of Local Conditions." *American Sociological Review* 62:94-112.

Petersen, Trond. 1991. "Time-Aggregation Bias in Continuous-Time Hazard-Rate Models." Pp. 263-90 in *Sociological Methodology*, edited by Peter V. Marsden. Cambridge, MA: Basil Blackwell.

Petersen, Trond and Kenneth W. Koput. 1992. "Time-Aggregation Bias in Hazard-Rate Models With Covariates." *Sociological Methods and Research* 21:25-51.

Rogers, Everett M. 1995. *The Diffusion of Innovations*. 4th. ed. New York: Free Press.

Soule, Sarah A. and Yvonne Zylan. 1997. "Runaway Train? The Diffusion of State-Level Reform in the ADC/AFDC Eligibility Requirements, 1950-1967." *American Journal of Sociology* 103:733-62.

Strang, David. 1991. "Adding Social Structure to Diffusion Models: An Event History Framework." *Sociological Methods and Research* 19:324-53.

Strang, David and Sarah A. Soule. 1998. "Diffusion in Organizations and Social Movements: From Hybrid Corn to Poison Pills." *Annual Review of Sociology* 24:265-90.

Strang, David and Nancy Brandon Tuma. 1993. "Spatial and Temporal Heterogeneity in Diffusion." *American Journal of Sociology* 99:614-39.

Theil, Henri. 1971. *Principles of Econometrics*. New York: John Wiley.

Thornton, Patricia H. and Nancy Brandon Tuma. 1995. "The Problem of Boundaries in Contemporary Organizational Research." *Academy of Management Best Paper Proceedings*: 276-80.

Tuma, Nancy Brandon and Michael T. Hannan. 1978. "Approaches to the Censoring Problem in Analysis of Event Histories." Pp. 209-40 in *Sociological Methodology*, edited by Karl F. Schuessler. San Francisco: Jossey-Bass.

———. 1984. *Social Dynamics: Models and Methods*. Orlando, FL: Academic Press.

Wu, Lawrence L. 1996. "Inconsistent Estimation Under Left-Censoring." CDE Working Paper No. 96–26, Center for Demography and Ecology, University of Wisconsin at Madison.

Henrich R. Greve is an associate professor at the University of Tsukuba in Japan. His current research includes work on an event-history approach to the study of heterogeneous social influence processes (with Nancy Brandon Tuma) and on the effect of adaptive aspiration levels on organizational change. Recent publications include "The Effect of Change on Performance: Inertia and Regression Toward the Mean" (Administrative Science Quarterly, *1999) and "Organizational Ecology and Job Mobility"* (Social Forces, *2000, with Takako Fujiwara-Greve). He has a Ph.D. in business administration from Stanford University.*

Nancy Brandon Tuma is a professor of sociology at Stanford University. She introduced event-history analysis to the sociological community and has continued to extend it throughout her career. Her current research in sociological methodology focuses on an event-history approach to the study of heterogeneous social influence processes (with Henrich R. Greve). In addition, she is studying the transition from socialism in post-Soviet countries in general and social inequalities in these societies in particular. She recently published Modern Russia *(coauthored by Mikk Titma). She has a Ph.D. in sociology from Michigan State University.*

David Strang is an associate professor of sociology at Cornell University. His research interests are in the study of organizations, political sociology, and methods of dynamic analysis. Much of this research has involved the study of diffusion through event-history-based analysis, textual analysis of media-based discourse, and simulation of adoption and abandonment trajectories. Current work focuses on the spread of work innovations like quality circles, total quality management, and reengineering in the American business community. His Ph.D. degree is from the Sociology Department at Stanford University.