Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect

ANDREW ABBOTT and ANGELA TSAY

Sociological Methods Research 2000; 29; 3
DOI: 10.1177/0049124100029001001

SPECIAL SECTION ON SEQUENCE ANALYSIS

The authors review all known studies applying optimal matching or alignment (OM) techniques to social science sequence data. Issues of data, coding, temporality, cost setting/algorithm design, and analytic strategies are considered, and substantive findings are reviewed. The authors conclude that OM techniques have produced interesting results in a wide variety of areas, the most promising being studies of careers and of sequentially organized cultural artifacts.

Sequence Analysis and Optimal Matching Methods in Sociology

Review and Prospect

ANDREW ABBOTT
ANGELA TSAY
University of Chicago

1. INTRODUCTION

A wide variety of sociological questions take the form of what may be called sequence questions, questions about whether some process or series of events typically happens in a particular order. Such questions are central in the literature on careers of various kinds: occupational careers, criminal careers, organizational careers. They are also common in studies of the life course and its transitions, as well as in a wide variety of macrosociological fields where nouns such as *professionalization* and *modernization* provide shorthand summaries of

AUTHORS' NOTE: *This article was presented at the American Sociological Association Methodology Section's midwinter meeting in Chicago, April 6, 1998. We thank Larry Wu for providing helpful comments on the article. The present version, particularly its final section, owes much to his concerns.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 29 No. 1, August 2000 3-33

theories that organizations and nations develop following characteristic patterns. In all these fields, beyond the simple question of the existence of patterns, lie other questions about what influences those patterns have on other variables and, indeed, what other variables dictate the patterns that cases follow. Related to these questions, mathematically although not substantively, are questions about similar "orders" in social and cultural structures. The question of whether the various elements of a story follow the same pattern in all variants poses a career-like problem, although a story does not have the stochastically generated character of an occupational career. The same is true for the question of the characteristic arrangements of stores in strip malls or of neighborhoods along transit lines.

The term *sequence analysis* has come to be applied to a variety of methods in sociology that address these kinds of questions. All of them have the common property of taking as their input sequences of data rather than individual data points. In formal terms, the input data are ordered arrays. Usually these are one-dimensional, although they are not necessarily so. The identifying characteristic of sequence data is their ordered character, not their dimensionality.

The most familiar form of sequence data, in this sense, is time series data. Such data are of course the object of well-known methods, which treat them as generated by an underlying stochastic process. The common characteristic of sequence analyses is that they treat each data sequence as a whole rather than as stochastically generated from point to point. It is as if one were comparing many time series to one another as whole units.

There are a variety of sequence analytic methods at present, and undoubtedly more will emerge (for a review, see Abbott 1995b). Currently, two general types of sequence analyses are becoming common: event structure analysis, developed by Heise (1991), and optimal matching or alignment (OM) methods. This article reviews applications of the latter to social scientific data.

2. *HISTORY*

OM methods were developed for the rapid analysis of protein and DNA sequences. In biology, the characteristic task was to search a

large database for close matches to a particular sequence of interest, typically a recently discovered protein. These "template" applications far outnumbered the "multiple alignment" studies, in which a number of proteins or DNA sequences from various organisms would be analyzed ensemble to create a distance matrix that could then be clustered in order to ascertain patterns of descent. The reason was simple: The template problem was linear in the size of the data set whereas the multiple alignment problem was quadratic. Outside biology, too, the most common problems were template ones, which arose in computer science under the term *string editing*. These typically occur in error correction routines such as spell checkers and file comparison algorithms.

OM algorithms first appeared in the early 1970s. By 1980, there was a substantial body of research, reviewed by Sankoff and Kruskal (1983). By the 1990s, string algorithms (as they are now usually called) had become a standard course in computer science training programs.[1] The rapid proliferation and widespread use of string algorithms led to an important change in the way they were conceived. Early applications generally viewed the algorithms as mimicking actual processes: in the biological case, the processes of insertion, removal, mutation, transposition, and recombination. However, the algorithmic intractability of most of these processes led to an overwhelming emphasis on the simplest of them. As a result, OM algorithms are today conceived less as actual models for reality than as generalized pattern-search techniques. It is this general pattern-search capability that has led to their application in social science.

OM algorithms were first used in social science by Abbott and Forrest (1986). There have been a variety of applications since, and it is worthwhile to reflect on the utility of the algorithms and their potential for future development.

3. DEFINITIONS

OM algorithms generally work by defining simple algebras that permit the creation of metric distances between sequences. These simple algebras involve at a minimum the operations of replacement, insertion, and deletion, the latter two often known simply as indel. The

distance between one sequence and another is defined as the minimum combination of replacements and indels required to transform one of a pair of sequences into the other. Different replacements can be weighted differently in accordance with some theoretically driven scheme. Indels can be weighted linearly or, as is more common in the biological literature, assigned a single "gap cost" that may or may not be augmented by a smaller cost linear in the length of the inserted (or deleted) material. The indel weight can also vary with respect to replacement. One important variation in applications of OM algorithms in social science, then, is the setting of these various costs. However, there are some algorithms, generally those directed at much fainter types of regularities, that do not involve cost setting, but simply define weighted objective functions on the data that can be maximized to find shared patterns (see section 8).

Finding the minimum cost distance between two sequences, subject to given costs, is a nonobvious task. It is solved by dynamic programming, programming that handles a large computing problem by solving a set of smaller problems whose results depend "dynamically" on one another. In OM, that dependence takes the form of ordering the smaller problems recursively. The exact nature of the recursion depends on the elementary operations chosen for the underlying sequence algebra. Although these may potentially include such things as swaps of adjacent elements or even a reversal of some large subsequence, in practice the vast majority of alignment work in biology employs only substitution and indels. This reflects computational difficulty rather than biological reality; swap algorithms (e.g., CELLAR; see Wagner 1983) are extremely cumbersome.[2]

An application of OM typically proceeds in several steps. The data must first be coded into a set of sequences using a finite alphabet of states. A replacement cost matrix must then be defined on these states and a gap cost scheme chosen.[3] The algorithm is then applied, resulting in a matrix of distances between all pairs of sequences. This matrix itself must then be analyzed, typically with some form of dual-data reduction scheme such as cluster analysis or multidimensional scaling. In what follows, we shall review the strategies chosen at each of these steps by the various studies reviewed.

We have been able to locate 23 applications of OM in the general sociology literature, including articles, conference papers, and

dissertations. A number of other applications are in progress but at present are unavailable to us.

4. CHOICE OF DATA

The OM literature involves a bewildering variety of topics, from careers to daily life to national histories. Career analysis is the most common application. Chan (1995) and Halpin and Chan (1998) investigated careers at the class level and examined individuals' histories of class status over the life course. Abbott and Hrycak (1990), Carpenter (1996), and Stovel, Savage, and Bearman (1996) considered careers at the more detailed level of particular types of jobs. (The latter two cases work within an internal labor market, the former within a broader but still well-defined market structure.) Han and Moen (1999) combined these two levels into a three-track sequence of occupation, organization, and socioeconomic status. Blair-Loy (1999) also investigated career sequences that combine occupation and organization codes. Giuffre (1999) investigated careers in terms of structurally equivalent positions in a block-modeled network. Erzberger and Prein (1997) combined work and family sequences. Modell's (1997) analysis of patterns of communion attendance over the life course examined another type of "career." Still other analysts have looked at the careers not of individuals but of larger social structures. Stovel (1994) investigated the existence of historical patterns in rates of lynching in counties of the southern United States, whereas Abbott and DeViney (1992) considered the sequential onset of various welfare state programs in the major nonsocialist countries.

A number of authors have examined shorter sequential patterns. Wuerker (1996) studied sequences of types of service utilization among Los Angeles mental patients during a single year. Wilson (1998a, 1998b) considered microlevel sequences of activities within a single day among a random sample of Canadian women. Pentland et al. (1998) examined variety in work sequences at a reference library, a PC laboratory, and a travel agency. Sabherwal and Robey (1993, 1995) investigated development sequences for information systems in organizations, whereas Poole and Holmes (1995) studied sequences of group decision making in experimental groups.

Finally, a number of works have examined sequence patterns not in real time but in the formal order of cultural artifacts. Abbott and Forrest (1986) applied OM to figure sequences in dances, nesting that analysis within a larger question about historical patterns. Forrest and Abbott (1990) investigated patterns in folktales, and Abbott and Barman (1997) studied the rhetorical structure of sociology articles. Levitt and Nass (1989) examined chapter sequences in physics and sociology texts.

The broad topical range of these works confirms the many possibilities of the method. Questions of sequence regularity and pattern do indeed pervade the social sciences.

5. CODING AND TEMPORALITY

The types of data used have distinct effects on the types of coding. Coding is an important issue in OM for a number of reasons. The decision of how events are to be defined shapes the input irrevocably. Particularly where events are complex, decisions about how to group them are quite consequential. At the same time, the existence of weighted replacement and indel costs means that coding variations actually have less implication than it might seem. One analyst might lump events that another might split. But even an analyst who split events into subevents would mostly likely assign replacement costs between those subevents that were much lower than those between the larger events that such a splitter would share with a lumper. Thus, closeness information is not totally lost in the move from splitting to lumping.

A number of works have accepted organizationally determined codes. Thus, Wuerker (1996) used various officially recognized types of mental health services to code her "careers of utilization" as rank-ordered sequences. Carpenter (1996) used official job titles at the U.S. Department of Agriculture to code job histories, although he retained durational information. Others collapsed such official classifications (e.g., Stovel et al. 1996). Still others have used official numbers, but as rates. Stovel (1994) used numbers of lynchings per annum in southern U.S. counties. Modell (1997) used numbers of communions per year by individuals in a Swedish town. Finally, some writers

have used codings that are standard throughout the literature. Wilson (1998a, 1998b) used the time budget codes employed throughout Canadian studies of time allocation, whereas Chan (1995) and Halpin and Chan (1998) used versions of the Goldthorpe class scheme. Other authors have developed their own quite elaborate coding schemes. Pentland et al. (1998) used field-coded observation data, whereas Poole and Holmes (1995) applied an elaborate multilevel coding scheme to experimentally generated videotaped data. Giuffre (1999) produced her sequence elements by block modeling (for each year of her study) a 159×159 matrix of connections between photographers. Levitt and Nass (1989) had independent judges create a coding scheme for their textbook subjects, whereas Sabherwal and Robey (1993) derived their own coding scheme for innovation events by a process of progressive refinement, a process similar to that of Abbott and Barman (1997) and Forrest and Abbott (1990).

Forrest and Abbott (1990) conducted the only direct study of the impact of coding on OM results. They had five coders code the dance sequences they had analyzed in 1986 and examined the variation in the results. Even though the coders varied both in their particular codings and, especially, in the level of detail, the results were remarkably stable, both in the comparative results (using scaling and clustering) and in Monte Carlo significance testing of the resemblance of the actual distances produced.

A number of authors have created complex "events" for their sequences by cross-classifying a number of simple events. This is the only strategy available for dealing with multiple, parallel tracks of sequence information in the OM framework; it must be reduced somehow to the unilinear structure expected by the OM algorithms. Abbott and Hrycak (1990) combined codes for sphere of activity and particular job, whereas Abbott and DeViney (1992) coded welfare states in terms of which combination of the five basic welfare programs they possessed in any given year. Dijkstra and Taris (1995) coded individuals by combinations of household status (five types) across educational status (three types) across job status (three types). Blair-Loy (1999) combined a nine-category classification of types of jobs held by individuals with a four-category classification of the sizes of the firms within which they held those jobs. Han and Moen (1999) combined a job classification, an organizational classification, and a work

status code in their three-level sequences of work histories. These various combinatoric codings of events create serious complexities in the setting of replacement costs, as we shall see.

Related to issues of coding are those of temporality. The vast majority of career studies code their sequences with regular time intervals, usually single years. In such a case, a data sequence would consist of the type of job held each year for a number of years. A few studies (e.g., Wuerker 1996; Levitt and Nass 1989) used episodes (units of varying length) as their sequence unit. Both schemes allow repeats. Only Dijkstra and Taris (1995) used a purely rank-type order in which a code was entered only if it was different from the preceding code. Their application used a rather curious process of deleting "superfluous" events that has strange consequences for the distances observed (see Abbott 1995a). Abbott and Barman's (1997) subsequence application also used a rank-type order.

Abbott and Hrycak (1990) proposed several different temporal structurings for the same data. For example, their analysis reduced all careers to a single length and then aligned the *proportions* of the career spent in different kinds of jobs. Abbott and Hrycak also attempted a purely rank-based scheme and a scheme of logging durations to reduce the effect of long constant runs. Chan (1995) followed Abbott and Hrycak in attempting a "standardized temporality" model and found, as did Abbott and Hrycak, that the differences in results were not large. Only Stovel (1994) further developed the idea of varying temporalities. Interestingly, although Stovel's county lynching sequences showed no patterns under standard temporality, she hypothesized that lynchings might have important legacy effects through their retention in local memory and, hence, allowed each lynching to enjoy a finite, decaying presence for a few time periods after the one in which it occurred. Under this new temporality, strong patterns appeared.

In summary, OM has broad applicability in terms of data arrays. One obvious drawback of OM analysis—in comparison with, say, Heise's (1991) event structure analysis—is its unilinear nature. However, authors have to a certain extent evaded this problem by employing "combination events." They have also used flexible temporal schemes. Indeed, some have attempted several time schemes on one body of data and others have undertaken perturbation analyses of

coding and weighting. In general, none of these multiple analyses found substantial variation in results. Overall, authors do not consider the method to be excessively constraining with regard to the kinds of sequential data structures it can analyze or excessively vulnerable to the exact coding structures used for time and events.

6. REPLACEMENT AND INDEL COSTS

Having coded data, analysts must then set replacement and indel costs for the algorithms to use. Again, there has been a variety of choices. A few writers (Dijkstra and Taris 1995; Pentland et al. 1998) simply set all substitutions and indels to a single figure on the grounds that they lacked any theoretical reason for doing otherwise. Most authors, however, set distinct substitution and indel costs.

Most authors have specified replacement or substitution costs with a simple matrix. Some categorize the matrix elements in bands so that there are in fact only two or three different replacement costs (e.g., Levitt and Nass 1989; Chan 1995). Usually, these bands embody a linear ordering of some sort on the matrix elements, as in Giuffre's (1999) ordering of different kinds of blocks in a network. Some assign substitution costs on the basis of some known linear property of the categories (e.g., Carpenter [1996] uses job salary). Others have used theoretically generated costs. Still others have derived costs from rater rankings (Abbott and Forrest 1986; Forrest and Abbott 1990), usually organized into hierarchies. A number have used hierarchical cost structures in which there are major costs between large branches and minor costs between small ones. (The Beldings software used for several of these studies actually takes its cost input in this fashion [see Abbott and Hrycak 1990]. A similar cost structure was used in the OM section of Abbott and Barman [1997], although with different software. For further information on software systems, see the appendix.) Those who have used cross-classification of simple events to develop more complex "sequence events" have had to derive highly differentiated cost schemes (e.g., Abbott and Hrycak 1990; Abbott and DeViney 1992; Stovel et al. 1996). Often, these studies have combined several types of information. Thus, in the case of Abbott and Hrycak (1990), the substitution costs combined information on job

differences, sphere-of-activity differences, and transition information within both realms. Stovel et al. (1996) combined information on bank branch differences with information on job differences using transition information as a measure of distance for both kinds of differences.

OM analysis clearly needs to reflect on substitution costs. According to Stovel et al. (1996), "The assignment of transformation costs haunts all optimal matching analyses" (p. 394). Individual costs are of course one-dimensional, as is alignment itself. But the whole cost structure may embody several dimensions of difference. During the process of coercing multidimensional structures into unidimensional ones, information inevitably will be lost. We have yet to see studies attempting several substantially different cost schemes on cross-classification event data, even though the relative importance of various differences used to constitute a single replacement cost might differ with the context.

Perhaps more important, most studies do not investigate the effects of perturbation in costs on the results. Without more of this perturbation analysis, it will be difficult to assess the dependence of OM on particular cost schemes or, indeed, to develop criteria for judging any particular cost scheme effectively. Indel costs are also problematic. The original applications of OM in social science were done with what we can now see were relatively high indel costs. Abbott's early applications all set the indel cost at a value equal to the largest substitution cost plus the difference between the largest and the next largest (as in the original Beldings programs; see the appendix). A number of authors followed Abbott in this high value, usually setting indels equal to at least the largest substitution cost (e.g., Stovel et al. 1996; Pentland et al. 1998). However, some analysts have attempted several different insertion costs (e.g., Levitt and Nass 1989; Poole and Holmes 1995; Chan 1995) without noting serious variation in results. (Chan [1995] actually used a fixed gap weight plus a small amount linear in the length of the insertion, but the fixed gap weight was relatively high.)

However, S. Gauvreau (personal communication, 1994) pointed out that such high-cost indels coerce the algorithms in important ways. In particular, if sequences are of equal length and indels are set to any cost greater than half the largest substitution cost, indels will *never* be used, since it takes two of them to take the place of a substitution. Where sequences are unequal in length, indel costs of this size

would prevent the algorithms from using any more indels than exactly enough to offset that difference in length. As a result, it became clear that for most applications in which serious alignment of similar portions of sequences was desired, indels should be set much lower. To determine appropriate levels, it is necessary to actually watch the effect of changing the indel level on a number of alignment pairs. Some software packages have such a facility. For example, the EXPLORE module of Abbott's OPTIMIZE program allows visual inspection of the alignment of any pair of sequences, allowing the user to shift the indel cost and watch the effect on the alignment. (Some biological packages also have this facility.) Using this module, Abbott and colleagues (see Blair-Loy 1999; Carpenter 1996) ascertained that indel costs in the vicinity of 0.1 times the largest substitution cost tend to pick up the sequence regularities that appeared to be substantively interesting.

The area of cost setting thus has seen some real diversity. The majority of studies have used fairly simple substitution costs and fairly high indel costs. Although these have produced interpretable results (see below), there remains some question about their theoretical reasonableness. Cost analysis will be important in the development of OM analyses.

With respect to the issue of sequence length, we should mention here the corrections used to deal with the fact that variation in sequence length means that some pairs of sequences have a greater potential distance between them than do others. Abbott and many others have dealt with this by the expedient of dividing the ultimate cost of transformation by the length of the longer sequence of the pair. (Abbott's OPTIMIZE program does this automatically, as do the Beldings program and Stovel's SAS programs.) Dijkstra and Taris (1995) arbitrarily "reduced" all sequences to equal length. We should also note that in applications where all sequences are the same length by design (e.g., Giuffre 1999), the length issue of course does not arise.

7. ANALYTIC STRATEGIES

Once costs have been set and sequences submitted to the algorithm, the resulting distance matrix must itself be analyzed. There is a

distinction here between those who have done simply a first-stage analysis such as scaling or clustering and those who have gone on to apply some further analytic procedure to the clustered/scaled results. Loosely, we could consider this a distinction between those who have used the methods purely descriptively and those who have gone on to apply some form of causal analysis.

A few analysts have generated groupings more or less inductively. Among these appear to be Modell (1997) and Carpenter (1996). Pentland et al. (1998) examined sequential variety in groupings that are established ex ante—a solidly theoretical strategy. A few authors have used scaling as a first-stage analytic strategy (Abbott and Forrest 1986; Abbott and DeViney 1992). More commonly, scaling is used as an adjunct to cluster analysis (Abbott and Hrycak 1990; Poole and Holmes 1995).

Beyond these few, by far the most common first-stage analytic strategy is cluster analysis. The usual variety of cluster methods has been attempted. Perhaps more to the point, given the reservations many sociologists have about cluster analysis, a number of methods of cluster validation have been attempted. Some analysts have relied on interpretability as a validation criterion (e.g., Chan 1995; Halpin and Chan 1998). Abbott's first OM work with clustering (Abbott and Hrycak 1990) used jackknife tests comparing within- and between-cluster distances, a strategy more or less followed by Sabherwal and Robey (1993, 1995), who also examined sharp changes in fusion coefficients, as did Poole and Holmes (1995), Wuerker (1996), and others. Han and Moen (1999) used an ANOVA-type validation strategy somewhat like Abbott and Hrycak's (1990) jackknives. Various authors (Abbott and Hrycak 1990; Stovel 1994; Sabherwal and Robey 1993) identified a real or theoretical typical sequence in each cluster. Abbott and Hrycak (1990) used this typical sequence to test "quick identification" for unclassified sequences, offering a method for rapid classification of large sequence databases after preliminary clustering of samples from them. Stovel et al. (1996) also used such a quick classification procedure.

There has thus been some attention to the direct validation of cluster results. This validation has also come indirectly from the inspection of linear variables correlated with cluster membership, a procedure commonly used even if further analysis was not built on those

variables. Sabherwal and Robey (1993), Carpenter (1996), Han and Moen (1999), Wuerker (1996), Giuffre (1999), and Halpin and Chan (1998) all used such inspection to some extent.

The validation problem remains a central one for OM analyses, as it is for classificatory and descriptive methods generally. Only a handful of the studies use external criteria or probabilistic validation, the majority relying on theoretical interpretability or careful reading of fusion coefficients. Future studies will need to incorporate serious validation procedures more routinely.

Although sometimes used for validation, linear or categorical variables attached to cases may also be used in further analysis, with sequence clusters being used as either independent or dependent variables, a strategy discussed in Abbott (1990b). In such cases, the proof of the classificatory pudding comes in the explanatory eating.

Abbott and DeViney (1992) inaugurated this independent variable analysis of matchings by investigating the effects of various country attributes on the location of a country's welfare sequence in the two-dimensional scaled space of welfare sequences. Thus, sequence was the dependent variable. Results, however, were either negligible or counterintuitive. Poole and Holmes (1995), in contrast, found fairly clear independent effects of sequence type on three variables of their decision sequences: consensus change, decision quality, and decision satisfaction. Carpenter (1996) also found some independent variable effects, in his case effects of sequence type on total duration of employment in the U.S. Department of Agriculture. Han and Moen (1999) found strong effects of career pathway type on timing of retirement (the latter measured as receiving a pension and, thus, independent of the in/out of the labor force element of the sequences themselves). Thus, there have been some successful searches for independent effects of sequence, but no success with sequence as a dependent variable. This last point may reflect the difficulty of turning sequence type into a variable, which can be done either categorically or via a data reduction method such as the scaling used by Abbott and DeViney (1992). But we still lack substantial findings.

In summary, the chief analytic strategy with OM has been cluster analysis, usually with some form of direct validation. An important development here has been the use of quick classification schemes to overcome the inherent limitation of multiple alignment type

approaches, with their quadratic calculation loads. Because clustering programs have themselves become more powerful—handling data sets in the thousands—such classification schemes have been less necessary, but they still permit the analysis of truly massive data sets.

A few analysts have gone beyond clustering to look for independent effects of sequences on various outcomes and have found effects of varying sizes. None of these analyses produces the kinds of decisive effects that would encourage the use of the methods. But the descriptive studies produce reasonably interpretable results, and the explanatory studies produce modest effects. The dangers with a technique as descriptively oriented as OM lie in the possibility of substantial interpretation of artifactual results. The literature to this point apparently does not suffer from this difficulty. Authors have been careful with the validation problem, and the substantive results are encouraging, if not overwhelming.

8. A NOTE ON SUBSEQUENCES

Often in sequence analysis, we are interested not in the resemblance of whole sequences but, rather, in the possibility of common subsequences. The various works on turning points (see Abbott 1997b) are a case in point. Here, we expect to find, in each sequence, some short pattern common to most sequences. There might be insertions in this subsequence or random replacements in this or that version of it. But still we expect a common subsequence, albeit with possible damage. Until recently, this problem was computationally difficult. However, an algorithm due to Lawrence et al. (1993) provides a solution, finding for any group of sequences the optimal common subsequence across the group. This optimal common subsequence need not appear in all sequences and need not appear in total in any single sequence. Indeed, it can start at points that vary widely across the sequences. The algorithm works via a Gibbs sampler embedded in an optimization structure. A Markov chain is generated on the space of all possible subsequence combinations using conditional probability information on transitions from one such combination to the next in such a way as to maximize a particular objective function defined on the combination.

A sociological illustration—applying the algorithm to the problem of finding subsequences in the rhetorical patterns of sociological articles—is provided in Abbott and Barman (1997). The algorithm locates a regularity that appears perfectly only in a small fraction of the data but that recurs in part throughout it. (There *do* turn out to be short rhetorical patterns that recur across sociological articles.) The results were validated via a Monte Carlo test in which final values of the objective function were calculated for hundreds of runs on random permutations of the data, providing a baseline distribution for the value observed on the unpermuted data. The results are quite striking and completely unachievable via direct methods.

9. SUBSTANTIVE RESULTS

We consider the substantive results of these analyses in the same order as we took up their subjects in section 4. We begin, then, with studies of individual careers. Chan (1995) found four distinctive job careers in a small sample ($N = 37$) of Hong Kong residents who have moved into the service class. These results persist under various transformations of cost and temporality. Chan argued that the congruence of standardized and real-time results indicates the absence of tournaments in mobility. Chan also found some distinct socioeconomic correlates of these careers.

Halpin and Chan (1998) undertook a much larger study, analyzing about 1,000 sequences each in two different data sets (the Irish Mobility Study and the British Household Panel Study). The data are retrospective class-level histories covering ages 15 to 35. Because all data are artificially set to be of equal length, Halpin and Chan felt they could afford to set the indel cost relatively high. Sequences were "padded" on the left with a "waiting to enter the labor market" state. Halpin and Chan were less concerned with documenting particular sequence clusters than they were with investigating the power of OM in working with large data sets. Their conclusion was that OM provides a very powerful first cut at the data, which must then be supplemented by more traditional, narrowly focused methods. Period effects, cohort effects, and specific sequence effects all became evident in the OM

results, providing a useful guide for how and, especially, where to attack data in more detail with standard methods.

The studies on job careers succeed to varying degrees. Abbott and Hrycak's (1990) analysis of several 100-person samples of German musicians in the eighteenth century establishes the existence of one major career ladder, thereby demonstrating that the entire musical labor market had at least some of the characteristics of an internal labor market. Individuals who reached the topmost positions had generally done so via a limited set of routes. Other ladders, however, were shorter and less clear. Studying 1,000 careers, Carpenter (1996) demonstrated that an integrated system of career trajectories, personal networks, and careful selection played a central role in building the comprehensive organizational culture that made the U.S. Department of Agriculture fundamentally different from other federal agencies. Whereas prior authors had emphasized the origin of this culture in various forms of political control, Carpenter demonstrated the role career trajectories played in knitting the structural basis of culture and commitment in the agency. In contrast, Stovel et al. (1996) emphasized change in career trajectories over time, demonstrating the emergence (and deliberate creation) of a new achievement-based career at Lloyds Bank in three broad cohorts whose careers began in 1890-1910, 1910-1925, and 1925-1934. More important, however, the study demonstrates that the career experiences of cohorts cannot be effectively inferred from instantaneous promotion rates. The promotion-rates analysis implies that mobility was greatest at the beginning and the end of the period investigated, but the OM analysis of actual careers shows that the actual career expectations of the first two cohorts had more in common with each other than either did with the later cohort. The interaction of bank expansion, individual choices, and career ladder structures produced individual life experiences quite different from those that would be expected by simply combining promotion chances with artificial cohort experiences.

Han and Moen's (1999) analysis of 500 retirees found career trajectory to be the most powerful independent predictor of retirement timing and expected retirement plans, as well as a crucial mediating variable for the effects of other independent variables such as gender. Trajectories even seem to affect, although weakly, postretirement employment, a variable otherwise unaffected by all the predictors

investigated. Blair-Loy's (1999) more detailed analysis of a small sample (N = 56) of women currently in successful finance careers revealed a clear differentiation among three cohorts of careers: careers starting before 1969, careers starting from 1970 to 1973, and careers starting after 1974. Blair-Loy also differentiated several types of careers, from "corporate climbers" to "big fish in small ponds" to "movers and shakers" and "entrepreneurs." Erzberger and Prein (1997) examined the work/family careers of 129 men and women and discovered considerably more commonality across gender than they had expected. A conspicuously elegant analysis by Giuffre (1999) shows that photographers with particular trajectories in terms of network location receive vastly different levels of media attention. Giuffre turned to network theory for theoretical rationales, invoking both structural holes and weak ties as sources for this very strong finding.[4]

These studies of careers demonstrate the utility of OM as an exploratory strategy, finding types and patterns that can then be used as dependent or independent variables in further analysis. Halpin and Chan (1998) support OM as a general exploratory strategy. Moreover, Stovel et al. (1996) make the important point that OM can uncover regularities that although central to individuals' lives are invisible in most one-time aggregate measures. It is clear that to the extent that one focuses on the life course, whether in itself as a site of experience or for its effects on other social phenomena, OM is a useful, perhaps even a central, technique.

We turn now to investigations of career-type patterns of larger social structures. Abbott and DeViney's (1992) analysis of the historical evolution of 18 welfare states characterizes them by whether they possess in each year of their careers the five basic welfare programs: workers' compensation, unemployment insurance, old-age pensions, national health insurance, and family allowances. The OM analysis captures most of the variation in these sequences in a two-dimensional scaling whose dimensions reflect, for the most part, variation in the relative onset of health insurance and family allowances. A large battery of dependent variables—from "modernization" variables such as levels of agricultural production to "demand" variables such as size of socialist parliamentary delegation to "state-based" variables such as taxation levels—predict very little of this variation. Abbott and DeViney then turned to diffusion accounts for it, using classical

spatial autocorrelation techniques but finding little diffusion effect. Ultimately, they settled for a probabilistic model in which the program adoption sequences emerge from random draws on pan-European hazard rates for the various programs, suggesting a purely institutionalist, common-origin account.

Stovel's (1994) elegant analysis of lynchings in 395 southern U.S. counties over 49 years contains a number of findings. First, Stovel demonstrates that lynchings live in local memory; there are no strong sequence patterns unless one allows each lynching to contribute to county "lynching experience," not only in the year in which it occurs but also, up to a nonlinear decay, in subsequent years. Second, Stovel explored the pulsing pattern characteristic of lynchings. Even the most violent county does not have steady lynchings year in and year out. Rather, there are pulses of various types. Characterizing pulses in terms of intensity (numbers of lynchings), tempo (the length of one lynching cycle), and duration (the total number of cycles observed), Stovel found a number of fairly specific patterns. Most interesting, in the detailed case of the 226 lynchings in the state of Georgia, Stovel was able to show that the sequence patterns correspond very strongly to lynch "scripts" uncovered independently by an earlier analyst (Brundage 1993). Thus, the temporal pattern of lynchings over time is tied to the type of lynching (terrorist, private, etc.) that tends to predominate, suggesting a larger pattern.

The Abbott-DeViney and Stovel analyses suggest that in the larger context of transindividual actors, sequence patterns may be quite complex. This may reflect the greater degree of contingency and averaging that comes automatically with size, or it may reflect data constraints. But although the results here are tantalizing and interesting in many ways, they do not possess the clarity of the individual career studies.

The works on shorter patterns are quite diverse. Wuerker (1996) analyzed the utilization histories of 49 "revolving door" mental patients—patients admitted to a skid-row mental health center who had at least 25 service episodes. (Wuerker used only the first 25 service episodes; although sampled in one year, the patients often have records going back several decades.) Wuerker found patterns that in many ways reflect changes in availability of psychiatric services. She sought points for intervention in psychiatric careers, noting that

current intervention plans are based on long-term averages rather than sequence patterns. Hence, patients who are heavy users of (expensive) emergency services were targeted. But Wuerker was looking for, and thought she may have found, clarity of the results in the individual career studies.

Wilson's (1998a, 1998b) exploratory analysis of the daily behavior patterns of 30 Canadian women found clear patterns only when durational information was omitted. Wilson attributed this fact to the relatively short duration of anchoring events (cooking, eating, cleaning up) and their scatter throughout the day. In the "short-form" (no duration) version of the records, the women separate distinctly into "travelers" (whose activities involve much going about, whose eating sequences are irregular, and who turn out to be members of more complex households) and "nontravelers" (whose activities involve fewer midday shopping trips and who are far more likely to live alone). Wilson felt that this exploratory analysis could be improved with more attention to the details of alignment.

Pentland et al. (1998) conducted an exploratory analysis of interactions at a library reference desk (59 interactions), a PC lab (32 interactions), and a travel agency (47 interactions). They aimed to expand traditional concepts of variety in the workplace by augmenting concepts of task variety (variety of what can be done) with sequential variety (variety in the order in which those things actually get done). Their findings are strong. Sequential variety varied inversely with task variety. The librarians did many more things but tended to do them in a fixed order. The travel agents had a narrow range of tasks (booking hotels, rental cars, and air travel) but did the various subparts of those tasks in a bewildering variety of orders, depending on the contingencies that arose. The analysis thus directly challenged unidimensional accounts of routinization in service work.

Sabherwal and Robey (1993, 1995) discussed the implementation of information services development projects ($N = 53$), coded elaborately from hundreds of student interviews with developers. The first study established six basic types of projects. Two of these types were somewhat rationalized: a "textbook" approach in which development proceeds as the books say it should and a "minimalist" approach that is textbook-like but lacks clear definition. Two were strongly externally dependent: "outsourced" and "off the shelf." Two were driven by

performance problems: "problem-driven minimalist" and "in-house trial and error." In the second study, the same 53 projects were studied slightly differently. Following a more traditional "variance" approach to theorizing project development, the projects were clustered based on profiles of rates of participation of certain crucial actors in them. These new clusters (slightly different from those of the preceding article) were then sequence analyzed in an effort to seek correspondence between actor participation profiles and sequence patterns. The two analyses seemed to overlay nicely, encouraging the authors to speak of a reconciliation of differences between the variance and process views of development that they traced to theoretical work on innovation from the early 1980s.

The Poole and Holmes (1995) study is part of a long series of studies on decision-making groups done by Poole with various colleagues. The data here were 40 experimental groups whose interactions were videotaped and coded using Poole's elaborate "phase-mapping" system. The study investigated the effect of group decision-making support systems—in this case a set of software for brainstorming—on actual patterns of decision making. In this study, sequence was an intervening variable between the experimental conditions (running the software, using it "manually," and using no such decision-making support) and the outcomes (consensus change, perceived satisfaction with decision making, and perceived decision quality). The main result was that the experimental conditions had little effect on the outcomes. Surprisingly, the decision paths—the sequences themselves—did turn out to have some effect, although it was not easily accountable theoretically.

The impression conveyed by these diverse studies of short sequence patterns is one of real possibility. These studies nearly all have strong findings and nearly all find OM to be of great utility in parsing out variation in sequence patterns. It is also important to note that all of them look to traditions in which process theorizing has been long standing: small-group analysis, time-budget studies, and innovation process analysis. It is clear that the methods will experience wider use in such venues.

The studies of cultural sequences are mostly by Abbott with various collaborators. In the original study using OM in the social sciences, Abbott and Forrest (1986) compared change over time in the

sequence of figures in four folk dances in a Gloucestershire village, aiming to test Cecil Sharp's view of the constancy of dance traditions within particular villages. The Sharp theory was strongly disconfirmed. Later, Forrest and Abbott (1990) undertook an extensive reliability analysis of the original study's coding scheme but also added an analysis of a celebrated set of versions of a Native American story, the star husband tale ($N = 67$). The authors' OM analysis replicated almost perfectly Thompson's ([1953] 1965) original laborious classification of the tales into groups, adding to that classification a transitional group that Thompson had missed. Together, the two analyses suggest strong reliability and validity levels for OM on divergent data sets.

Abbott and Barman's (1997) study, like Levitt and Nass's (1989) study of textbooks, concerned social science publishing. Abbott and Barman studied 100 articles from the *American Journal of Sociology* over a 70-year period to look for evidence of the increasing rhetorical rigidity thought to characterize sociological articles. Evidence for that increase is slight but definite. Surprisingly, there appears to have been a transition from an earlier, somewhat rigid form to a later, somewhat more rigid form via a period of real confusion about rhetorical proprieties. In the subsequence portion of the paper, Abbott and Barman uncovered a particular rhetorical gesture—the data-methods-analysis subsequence—that does increasingly pervade sociological writing.

Finally, Levitt and Nass (1989) examined ordered lists of chapter subjects in 45 textbooks in physics and sociology, seeking from OM evidence of isomorphism in the topics and order of texts. Evidence is overwhelming. Moreover, homogeneity is greater in physics than in sociology, which Levitt and Nass regard as "less mature" and more "preparadigmatic." Interview data buttressed these conclusions. Levitt and Nass took this as strong evidence that textbook publishing was dominated by the institutionalists' isomorphic processes rather than by the "garbage can" decision making of earlier theorists.

In some ways, the cultural applications of OM have been the clearest of all OM work. They have produced interpretable, intuitive findings. Although most of these have been classificatory, they have also been used effectively to test changes in cultural sequences over time. Given the number of cultural artifacts in which sequence plays a crucial role, this area, too, seems ripe for exploration.

10. ISSUES AND FUTURE DIRECTIONS

OM methods thus seem to have broad substantive application and a certain degree of robustness. In this closing section, we discuss some general issues and possibilities. By the standards of workaday sociological methods, OM is decidedly heterodox, and it is useful to recapitulate some of the points at issue between the OM approach and the mainstream. In a commentary on this article, Wu (2000) notes the striking contrast between its sober tone and Abbott's (1995b) strong claims in a review of the broader area of sequence analysis. In fact, this contrast derives from the different levels at which the debate over methods has been joined. At the most general level, OM reflects a drift in social science toward thinking about social reality in terms of "events in contexts" rather than "entities with variable attributes." In this drift it has much fellow flotsam: historical sociology, narrative analysis in cultural sociology, game theory, network analysis, and many others. We should not confuse the overarching debate between the events and variables paradigms with the much more specific issue of how OM relates to orthodox methods for cognate problems.[5] Being an active polemicist at the more general level (e.g., Abbott 1983, 1988, 1992, 1997a) should not disqualify one for actually trying to do what the polemic promises. Indeed, it should rather oblige one to do so; sociological theory is littered with unredeemed promises.

At the more immediate level of methods themselves, the orthodox community has two real queries with regard to OM. The first of these is simple: Why bother with OM, since "we already do this"? The second is both more substantive and more worrisome: It concerns the "unreality" of OM as a model for temporal analysis.

These two problems actually arise from a single difference. In the most immediate sense, of course OM does things that orthodox methods cannot. It classifies sequences. But the orthodox reply is that OM classification is based on arbitrary algebras and therefore is not scientific, whereas orthodox classification—the classification that is implicitly made when a set of sequences is shown to be not inconsistent with a particular generating probability process—is based on a model of the situation and thus *is* scientific. In that sense, orthodoxy *does* classify sequences whereas OM only appears to.

The argument applies to any application of OM to real-time sequences. At issue is the "reality" of what OM does. In the early biological applications, as we have noted, the algebraic apparatus of OM's dynamic programming echoed the physical processes of mutation (substitution) and cleaving and insertion (indel) that actually happen during the physical replication of DNA and proteins. However, some biological events—such as short swaps or transpositions—are now widely ignored because of their computational intractability, their effects being found via algorithms using more tractable elementary operations. Similarly, in social science, mutation and cleaving make literal sense only in those applications studying sequences in social or cultural space, not time. One can change the fourth element of a ritual or insert three different kinds of stores at the end of a strip mall; these are mutations and insertions in linear sequences at a given time. But in studies of careers, for example, the idea of "mutation" of a particular element of a sequence seems odd; careers are generated "from left to right," in real time. It seems natural to think of them in the durational methods framework, as the realizations of probability processes generated moment by moment.

So the questions with regard to OM are, first, whether it is legitimate to use a classification method that is not based on a probability model of sequence generation and, second, why one would want to do so if one could. The "whether" question is less problematic once we reflect on scientific inquiry outside our own field. All sorts of sciences have used and still use classification methods that are descriptive rather than causal, methods that are based on arbitrarily chosen attributes rather than causal mechanisms. Rather, it is the notion of describing things by analyzing what causes them—of which Durkheim's celebrated petitio principii at the beginning of Book II of *Suicide* is so perfect an example—that is philosophically worrisome. The only scientifically legitimate way to test hypotheses is to refer them to a measuring of social reality that they do not themselves define. Only purely descriptive methods can produce such a measuring, and therefore classification methods not based on probability models are certainly not prima facie illegitimate. The philosophical challenge is more on the other side.

But of course that such methods are legitimate does not necessarily mean we should bother with them. The more important question is

why one would want to use OM when event history (EH) methods are available. A number of reasons appear to be obvious. First, EH methods do not handle whole sequences of diverse events. They do all right with one or two competing risks. They can combine two recurrent processes into whole sequence models, as Lillard (1993) did for marriage and childbearing. But EH methods really do not make any sense for career trajectories or subpatterns involving dozens of types of jobs. They might work for Stovel's lynchings, but hardly for Blair-Loy's finance careers. So the first reason for doing OM is that it is pretty good at a set of tasks that EH is pretty bad at.

Second, EH methods have their own burden of assumptions, a burden that grows apace as we move away from the simple duration problem toward competing risks, multiple outcomes, and repeated events. The most important of these assumptions involves the independence of observations, the foundation without which maximum likelihood estimation becomes impossible. The obvious multiplex nonindependence that infects any durational structure is in practice handled by arbitrary assumptions about errors. Thus, the image that EH models are free of arbitrary assumptions whereas OM methods are plagued by them is false. On the contrary, EH models routinely make assumptions that seem heroic to those outside the orthodox community.[6]

Third, the idea that careers are generated by a probabilistic process independent across individual units is itself only a hypothesis. When we model reality with a hypothesis, we cannot find those aspects of reality that are not specified in that hypothesis. By the nature of our modeling, unforeseen things become the error against which we test our own guesses about what is going on. To take the fertility example, suppose that fertility is a function of how many of one's friends are having children—suppose one conceives a child if at least two of one's friends have done so in the past two years. In such a system, the network structure of society will be by far the most important determinant of individual fertility, yet this structure cannot be fully modeled in an egocentric fashion, unit by unit.

The advantage of OM, in this context, is precisely that it makes no assumptions about what produces sequence regularity. OM would find this particular regularity because people in particular friendship networks would turn up in groupings of similar fertility careers. Of

course, the data requirements would be enormous and so on. But the point is a general one. By not making modeling assumptions, OM acts as a true description. It finds things we might like to explain, and it may well point the way toward explanation. Orthodox methods cannot find what they have allocated to error ex hypothesi.

Given this, one obvious utility of OM in the context of orthodox methods is its utility as a method of data reduction or data search. In many social scientific data situations, most of the possible events that could occur do not occur (see Abbott 1990a). The total state space has large empty regions. When state spaces are empty, classification and typology are a useful preliminary to analysis. An example is the typical EH-based analysis of the passage of this or that type of law across the American states. The orthodox approach is to treat this sparse data space as "really full." One treats "state-years at risk" as independent on the assumption that all the unmeasured historical continuities (i.e., events) that link up Massachusetts in 1909 with Massachusetts in 1910 can be handled by an error term for heterogeneity. This draconian assumption allows one to treat 50 state histories as if they were hundreds of individual cases of data. But it might very well turn out that far from there being hundreds of "independent trials" of an "instantaneous process" of compensation law passage, there are in fact only three or four typical sequential patterns of passage for such laws. This fact could not be found by the standard approach, but only by OM or other whole sequence methods.

These various considerations suggest that OM analysis is both legitimate and useful. At heart, the difference between OM and EH lies in their stances toward data. To use the standard terms from the philosophy of science, OM is about discovery whereas EH is about justification. OM aims to find regular social phenomena to think about. EH already presumes a fully formed model of society and social causality and wants to test that model on data. (For more on that general approach to explanation, see Abbott [1998].) The two views have different intentions.

In closing, we would like to mention two important areas for future development in OM work. One major area for development is validation. The various strategies developed so far do not add up to a comprehensive approach to validation. Although the OM literature would

be assisted by more widespread use of standard approaches to cluster validation, more than that is needed. Required are serious Monte Carlo and perturbation analyses of OM applications to simulated data. Only these will begin to give us a real feel for the underlying measure issues involved in sequence analysis via OM. The issue of validation raises more general questions about the probabilistic realities associated with OM and the consequent statistical behavior of matching distances under perturbations. These are subjects on which even the biological literature has been relatively quiet (e.g., see Vingron and Waterman 1994).[7]

A second major area of interest is the application of OM to continuous data, what Sankoff and Kruskal (1983) call "time-warping." In comparing several time series—for example, of wage changes in industries—it might be that all of them have the same general shape over time but that some go through parts of that shape faster than others. Or it might be that there is a specific subseries that all of a set of time series contain at some point within them. There are a variety of possible sequence regularities in time series data (both unidimensional and multidimensional) that could be discovered with warping methods.[8]

In summary, the current prospects of OM are reasonably good. There is not yet any single, decisive application—one that completely solves a major empirical question left untouched by standard methodology or that completely overthrows standard interpretations. But there is a modest and growing record of applications, both in areas widely studied by standard methods and outside them. More general use of the methods will undoubtedly solidify this record and suggest resolutions for current problem areas such as substitution costs.

APPENDIX

A variety of software packages has been used in the literature. Most early work (Abbott and Forrest 1986; Abbott and Hrycak 1990; Forrest and Abbott 1990; Abbott and DeViney 1992; Poole and Holmes 1995; Sabherwal and Robey 1993, 1995) was done with the Beldings Program Series, a set of VMS FORTRAN programs originally written for CDC computers by David Bradley of Long Beach State University in the mid 1970s. In the early 1990s, Abbott developed OPTIMIZE (written in C), a small package for multiple alignment of small data sets (Carpenter 1996; Blair-Loy 1999). It includes a module for visual inspection of alignments and is available on the Web. Stovel's SAS/IML program DISTANCE has been used by some working with larger data sets (Stovel 1994; Stovel et al. 1996; Han and Moen 1999). Some have used standard biological packages, which include a larger variety of algorithms (e.g., algorithms with affine gap costs, such as PILEUP [Chan 1995; Halpin and Chan 1998] and CLUSTAL [Wilson 1998a, 1998b]). CLUSTALG, a social-science-friendly version of CLUSTAL, has recently been developed by Andrew Harvey and others (contact andrew.harvey@stmarys.ca). SEQUENCE, developed by Dijkstra (Dijkstra 1994; Dijkstra and Taris 1995), implements a rather curious algorithm. (Its properties are discussed in Abbott 1995a.) Goetz Rohwer has implemented an alignment program as part of a suite of event history programs (TDA; contact rohwer@mpib-berlin.mpg.de). The LEA Gibbs sampling algorithm for subsequences is available (in compiled QuickBasic) from Abbott directly. At this point, then, the would-be user has a considerable number of alternatives.

NOTES

1. Gusfield (1997) is a comprehensive computer science text on string matching. The book has three sections. The first concerns the exact string matching problem, typically the template problem of finding out whether a particular string matches part of some larger string or strings. The second section concerns suffix trees, a preprocessing algorithm designed to speed large-scale template problems for exact matching. Suffix trees in turn permit mismatching in compared sequences, which leads to the book's third main section, on inexact matching, alignment, and dynamic programming. Multiple alignment problems are discussed throughout this section. Social science applications have, to our knowledge, all employed dynamic programming. Gusfield's book contains an extraordinary variety of algorithms, which will greatly expand the social science OM literature.

2. Because our main aim here is to review particular studies, we cannot examine the details of OM algorithms. The standard texts are Sankoff and Kruskal (1983) and Gusfield (1997). Readers who may wonder (as one reviewer did) "where are the equations" should note that the task these algorithms carry out is not estimating parameters of equations but, rather, finding the minimum cost procedure for turning one sequence into another subject to given cost constraints. This is essentially a combinatoric task, solvable by recursive procedures discussed in the sources given. It should also be noted that the issue of the "reality" of OM's procedures—that is, the de-

gree to which they are models rather than simple analytic procedures—is an important one. We consider it in some detail in the final section.

3. Replacing element *a* in sequence 1 by element *b* to help change sequence 1 into sequence 2 adds a certain amount—set ahead of time—to the total cost of transformation. The matrix of these replacement costs (always taken to be symmetric) is the replacement cost matrix and must be set up before analysis. The gap cost scheme involves deciding the cost to be added for an insertion or deletion. As noted, this must be scaled relative to the replacement costs, and may be set up either as a single cost for each indel or as a fixed cost for any "gap" plus a penalty linear in the gap length.

4. We omit from this review Modell's (1997) brief exploratory use of OM in his study of patterns of communion attendance. Modell actually ends up doing an enumerational analysis rather than an OM one.

5. The drift toward events and contexts is even characteristic of the natural sciences, where analytically derived methods are in many places challenged by algorithmic and stochastic ones. Maximum likelihood estimation—the heart of orthodox methods for durational and sequence data—looks like a simple piece of calculus on paper, but when computers get down to the work of optimizing complex functions these days, they do not just calculate grad(l) and head downhill. They use simulated annealing or other probabilistic structures to wander around the solution space in a way that is much less systematic but ultimately much more certain. It is striking that some of the major workers in duration models use Monte Carlo techniques extensively (e.g., Heckman and Walker 1990).

6. See Lillard (1993:193) for an example of these assumptions, for example, the strong and to us unpersuasive assumptions about the identification of various independent factors shaping conception. The issue of unobservables and their degree of serial correlation plagues all event history (EH) models. (For more philosophical, as opposed to mathematical, assumptions of the orthodox view, see Abbott [1988]; for EH in particular, see Abbott [1990a]. Abbott is recasting certain of the arguments of the latter in a forthcoming study.) While the EH theory of the temporality of the social process is undoubtedly correct, the "deep past influence" argued in Abbott (1990a:146-47) does exist because past structural phenomena ("large events") encode themselves in existing social structure (where they have the practical effect of destroying intercase independence and producing linear dependence among predictor variables). Thus, the correctness of the EH view of temporality—the world really is Markovian—does not allow EH to escape from the impact of the deep past (see Abbott 1998). For equivalent arguments from the viewpoint of an economist, see the essays in Shackle (1990).

7. The lack of serious perturbation analyses makes us less able to judge the stability and reliability of the various analyses reported here than we otherwise would be. To provide a more pointed judgment of the quality of these studies requires a clearer sense of what levels of stability we should expect or require. It may well be, as noted earlier, that applications of OM should routinely include data on perturbation analysis of some sort. The measures reported in Forrest and Abbott (1990) are examples, although they rest on actual multiple codings of the raw data rather than on random perturbation.

8. In unpublished work, the senior author has found that standard algorithms do quite well in distinguishing certain kinds of continuous data. In particular, if we create three sets of regression lines originating at a single point, each set with its own slightly different slope, OM can distinguish the three sets even in the presence of quite strong random variation about the lines. With sine curves and similar variation, however, OM does less well, although it has not yet been combined with the obvious "operation" of phase shifting.

REFERENCES

Abbott, A. 1983. "Sequences of Social Events." *Historical Methods* 16:129-47.

———. 1988. "Transcending General Linear Reality." *Sociological Theory* 6:169-86.

———. 1990a. "Conceptions of Time and Events in Social Science Methods." *Social Science History* 23:140-50.

———. 1990b. "A Primer on Sequence Methods." *Organization Science* 1:373-92.

———. 1992. "What Do Cases Do?" Pp. 53-82 in *What Is a Case?* edited by C. Ragin and H. Becker. Cambridge: Cambridge University Press.

———. 1995a. "Comment on Dijkstra and Taris: The Scope of Alignment Methods." *Sociological Methods & Research* 24:232-43.

———. 1995b. "Sequence Analysis." *Annual Review of Sociology* 21:93-113.

———. 1997a. "Of Time and Place." *Social Forces* 75:1149-82.

———. 1997b. "On the Concept of Turning Point." *Comparative Social Research* 16:89-109.

———. 1998. "The Causal Devolution." *Sociological Methods & Research* 27:148-81.

———. 1999. "Temporality and Process in Social Life." Pp. 28-61 in *Social Time and Social Change*, edited by F. Engelstad and R. Kalleberg. Oslo: Scandinavian University Press.

Abbott, A. and E. Barman. 1997. "Sequence Comparison via Alignment and Gibbs Sampling." *Sociological Methodology* 27:47-87.

Abbott, A. and S. DeViney. 1992. "The Welfare State as Transnational Event." *Social Science History* 16:245-74.

Abbott, A. and J. Forrest. 1986. "Optimal Matching Methods for Historical Data." *Journal of Interdisciplinary History* 16:473-96.

Abbott, A. and A. Hrycak. 1990. "Measuring Resemblance in Social Sequences." *American Journal of Sociology* 96:144-85.

Blair-Loy, M. 1999. "Career Patterns of Executive Women in Finance." *American Journal of Sociology* 104:1346-97.

Brundage, W. F. 1993. *Lynching in the New South*. Urbana: University of Illinois Press.

Carpenter, D. 1996. "Corporate Identity and Administrative Capacity in Executive Departments." Unpublished Ph.D. dissertation, University of Chicago.

Chan, T.-W. 1995. "Optimal Matching Analysis." *Work and Occupations* 22:467-90.

Dijkstra, W. 1994. "Sequence: A Program for Analyzing Sequential Data." *Bulletin de Methodologie Sociologique* 43:134-42.

Dijkstra, W. and T. Taris. 1995. "Measuring the Agreement Between Sequences." *Sociological Methods & Research* 24:214-31.

Erzberger, C. and G. Prein. 1997. "Optimal-Matching-Technik." ZUMA-Nachrichten No. 40, Mannheim.

Forrest, J. and A. Abbott. 1990. "The Optimal Matching Method for Anthropological Data." *Journal of Quantitative Anthropology* 2:151-70.

Giuffre, K. 1999. "Sandpiles of Opportunity." *Social Forces* 77:815-32.

Gusfield, D. 1997. *Algorithms on Strings, Trees, and Sequences*. Cambridge: Cambridge University Press.

Halpin, B. and T.-W. Chan. 1998. "Class Careers as Sequences." *European Sociological Review* 14:111-30.

Han, S.-K. and P. Moen. 1999. "Clocking Out." *American Journal of Sociology* 105:191-236.

Heckman, J. J. and J. R. Walker. 1990. "The Relationship Between Wages and Income and the Timing and Spacing of Births." *Econometrica* 58:1411-41.

Heise, D. 1991. "Event Structure Analysis." Pp. 136-63 in *Using Computers in Qualitative Research*, edited by N. Fielding and R. Lee. Newbury Park, CA: Sage.

Lawrence, C. E., S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wooton. 1993. "Detecting Subtle Sequence Signals." *Science* 262:208-14.

Levitt, B. and C. Nass. 1989. "The Lid on the Garbage Can." *Administrative Science Quarterly* 34:190-207.

Lillard, L. A. 1993. "Simultaneous Equations for Hazards." *Journal of Econometrics* 56:189-217.

Modell, J. 1997. "The Representation of Human Lives in Social Science and Social History." Unpublished manuscript, Carnegie Mellon University.

Pentland, B. T., M. Roldan, A. A. Shabana, L. L. Soe, and S. G. Ward. 1998. "Lexical and Sequential Variety in Organizational Processes." Unpublished manuscript, Michigan State University.

Poole, M. S. and M. E. Holmes. 1995. "Decision Development in Computer Assisted Group Decision-Making." *Human Communication Research* 22:90-127.

Sabherwal, R. and D. Robey. 1993. "An Empirical Taxonomy of Implementation Processes Based on Sequences of Events in Information System Development." *Organization Science* 4:548-76.

———. 1995. "Reconciling Variance and Process Strategies for Studying Information Systems Development." *Information Systems Research* 6:303-27.

Sankoff, D. and J. B. Kruskal, eds. 1983. *Time Warps, String Edits, and Macromolecules*. Reading, MA: Addison-Wesley.

Shackle, G.L.S. 1990. *Time, Expectations, and Uncertainty in Economics*. Edited by J. L. Ford. Brookfield, VT: Elgar.

Stovel, K. 1994. "The Structure of Lynching." Paper presented at the Social Science History Association, October, Atlanta, GA.

Stovel, K., M. Savage, and P. Bearman. 1996. "Ascription Into Achievement." *American Journal of Sociology* 102:358-99.

Thompson, S. [1953] 1965. "The Star Husband Tale." Pp. 414-59 in *The Study of Folklore*, edited by A. Dundes. Englewood Cliffs, NJ: Prentice Hall.

Vingron, M. and M. S. Waterman. 1994. "Sequence Alignment and Penalty Choice." *Journal of Molecular Biology* 235:1-12.

Wagner, R. 1983. "On the Complexity of the Extended String-to-String Correction Problem." Pp. 215-35 in *Time Warps, String Edits, and Macromolecules*, edited by D. Sankoff and J. B. Kruskal. Reading, MA: Addison-Wesley.

Wilson, W. C. 1998a. "Activity Pattern Analysis by Means of Sequence Alignment Methods." *Environment and Planning A* 30:1017-38.

Wilson, W. C. 1998b. "Analysis of Travel Behaviour Using Sequence Alignment Methods." *Transportation Research Record*, No. 1645, 52 pp.

Wu, L. 2000. "Some Comments on 'Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect.' " *Sociological Methods & Research* 29:41-64.

Wuerker, A. K. 1996. "The Changing Careers of Patients With Chronic Mental Illness." *Journal of Health Administration* 23:458-70.

Andrew Abbott is Ralph Lewis Professor of Sociology at the University of Chicago. He is the author of The System of Professions, *a study of the division of labor. He has also written extensively on the problem of temporality in social science and developed novel methods for sequence data. His new book,* Department and Discipline, *is a study of the Chicago sociology department in relation to the larger discipline of sociology. He is completing a book titled* Chaos of Disciplines, *a theoretical and empirical analysis of knowledge change in social science and, more broadly, of self-similar social structures. Abbott's papers on temporality are collected in* Time Matters, *forthcoming from the University of Chicago Press.*

Angela Tsay is a graduate student in the Department of Sociology at the University of Chicago. She is currently the director of customer experience at zZounds.com, a division of HarmonyCentral.com Inc. Her interests lie in the sociology of emotions and of academic disciplines. She has studied changes in patterns of academic recommendations over time.