SEQUENCE COMPARISON VIA ALIGNMENT AND GIBBS SAMPLING: A FORMAL ANALYSIS OF THE EMERGENCE OF THE MODERN SOCIOLOGICAL ARTICLE

Andrew Abbott*
Emily Barman*

Various substantive literatures in sociology seek small regularities in sequences: turning points in the life course, catalytic moments in organizational change, sharp turns in occupational trajectories, and the like. Commonly these are turning points, but they may also be simple local patterns. This paper reports a method for discovering such regularities even when they are quite faint, applying that method to rhetorical regularities in sociological articles. The paper begins by analyzing the overall sequence structure of such articles and then gives a basic introduction to Gibbs sampling, one member of the broader class of Markov chain Monte Carlo (MCMC) methods. It then reports an algorithm employing Gibbs sampling to find local sequence regularities and applies that algorithm to demonstrate the subsequence regularities present in sociological articles. Substantively, the paper shows that the rhetorical structure of sociological articles changed from one pattern to another in the period 1895–1965 and that certain faint but standard rhetori-

*University of Chicago

*cal subsequences became characteristic of articles in the later pe-
riod. Methodologically, it introduces a broad class of methods that
provide effective approaches to a number of previously intractable
statistical questions.*

Sociological analysts often want to detect patterns in sequence data. Life
cycle theories, natural history models, and career concepts all hinge on the
notion of regular patterns in successive events. Methods for analyzing such
data, however, are not widespread. In a set of empirical papers, the senior
author has advocated the use of optimal alignment methods for categori-
zation of sequence regularities (see Abbott 1992, 1995a).

However, alignment methods work poorly with subsequence
problems. If our aim is to find a common short pattern within each ver-
sion of a longer sequence of events—the kind of pattern suggested by
the idea of "turning points" in careers, for example—global alignment
methods are too unfocused. In this paper we apply a new algorithm
developed for subsequence analysis by Lawrence and others (1993).
The algorithm uses Gibbs sampling, a data resampling strategy in the
tradition of the jackknife, the bootstrap, and other data augmentation
techniques.

At the same time, the paper aims to develop and analyze a particular
empirical question. A recent literature in the sociology of science has ex-
amined the rhetorical production of scientific texts. We follow this textual
tradition but apply formal sequence analysis methods to study the rhetor-
ical structure—in simple terms, the fundamental outline—of sociological
articles. The basic questions are whether there have been standard patterns
for articles and whether those patterns have changed.

Our multiple tasks impose a complex structure on the present arti-
cle. We begin with a necessarily brief discussion of the substantive prob-
lem. This leads naturally into a discussion of data, sampling, and coding.
We then introduce optimal alignment as a sequence analysis methodology.
Since this method has been discussed elsewhere, our discussion here merely
recapitulates, laying the groundwork for the later subsequence analysis.
Having presented the basic results of the optimal alignment analysis, we
then examine rhetorical subsequences in articles. Here we introduce Gibbs
sampling, first in general, then as a specific aid in analysis of subsequences.
The paper closes with a brief discussion of subsequence results and the
potentialities of methods presented here.

1. THE RHETORICAL STRUCTURE OF ARTICLES

The first literary analyst of social science, to our knowledge, was Joseph Gusfield, whose literary dissection of studies of alcohol abuse first began appearing in the 1970s. Gusfield's work sprang from that of Kenneth Burke (see Gusfield 1981), emphasizing a broad rhetorical analysis of the strategies of science for representing itself as authoritative. In the years since Gusfield wrote, several others have pursued his lead. McCloskey (1985) has analysed texts in economics. Myers (1990) discussed the rhetorical "taming" of articles in biology. Bazerman (1988) comes closest to the task we undertake here, examining the rhetorical structure imposed on psychology by the APA style manual and the detailed rhetorical patterns of articles in one volume of the *American Political Science Review*.[1]

While the rhetorical tradition has effectively reshaped our thinking about scientific writing, it has not empirically examined long-term changes in rhetorical patterns. Yet what persuades in one generation may not persuade in another. Indeed, the very relation of science to the rest of society has changed during the social sciences' existence, a change that in turn might be expected to change patterns of argument. Moreover, we expect changes in rhetorical patterns as disciplines themselves change and develop, growing now more, now less authoritative relative to competing disciplines.

There is thus a variety of mechanisms predicting changes in the rhetorical patterns of scientific argument. One particular generalization about that change is a common belief among sociologists: that over the years sociological articles have become more and more subject to rigid rhetorical conventions. This common opinion arises in our conceptions about disciplinary history. Most of us believe that the modern quantitative sociological article first appeared with the present tradition of quantitative sociology in the 1930s and that it became a formal, conventional structure by about 1965–1970, complete with theory section, literature review, presentation of data, discussion of methods, recounting of results, and conclusion. But there is little evidence of this trend other than some general

---

[1]O'Neill (1981) contains a good review of rhetorical studies of science as of that date. Kircz (1991) has analyzed article argumentation from the point of view of coding and informational retrieval. Clyne (1987) is an elegant comparison of the internal organization of English as opposed to German academic texts. Connors (1985) offers an actual history of explanatory rhetoric. Mullins et al. (1988) present a general plea for detailed analysis of articles.

statements about personal experience, all of which confound personal with disciplinary aging.[2]

In this paper, we formally evaluate the proposition that sociological articles have become more and more rhetorically fixed. We proceed by developing a lexicon of article elements (short "units" of sociological writing), coding a selected series of articles in terms of those elements, and then analyzing the results using sequence methods. If the "converging rhetoric" generalization is correct, we should find more recent articles to be more narrowly similar in structure at both global and specific levels. The precise nature of that narrow similarity should perhaps be left open. Next to nothing is known about this empirical generalization of convergence, so it would be premature to specify an expectation. But the overall empirical regularity we expect seems clear: convergence on a standard form.

2. DATA, SAMPLING, AND CODING

We propose to examine the question of convergence using a single journal, a stratified sample, and a formal method for coding. In this section we discuss these choices.

---

[2]This statement is common enough in speech but hard to document on paper. There is much informal data on the disappearance of unusual work from the journals. As one recent member of the AJS editorial staff put it to the senior author, "most of what we publish is extremely competent and extremely boring." Complaints on this score are old. Writing of the ASR, Page (1981) reports "in the fifties, according to many of its critics, the *Review* was theoretically thin, overloaded with small-scale research reports, and neglectful of important social and sociological issues—charges with which I agreed." Page turned to solicitation to get papers from scholars like Bierstedt, Goode, Hughes, Kroeber, Merton, Moore, Parsons, and Riesman.

There is some evidence on changes in methodology. See Smalley (1981), for example, whose data bears also on the trends in section 2.3.3, as well as McCartney (1970) and Bakanic, McPhail, and Simon (1987). Certainly changes in editorial processes might have produced increasing homogenization. Thus, Teevan (1980) had a large panel of blind judges rate published papers and found that ratings vary more within journals than between them, even though he used six general sociology journals with widely varying ratings on the Glenn (1971) prestige scale. This result suggests one dimension of homogenization. Another indication, if not of homogenization then of processes that could easily produce it, is the emergence of the "revise and resubmit" judgment, a judgment that has become very common in sociology journals in recent decades. Bakanic, McPhail, and Simon report 28 percent of articles were given R&R status on first review at ASR in 1977–1982. R&R could easily be a euphemism for a process enforcing standardization, although on another interpretation, the rise of extensive formal revisions would not be producing sterile homogenization, but rather "high standards."

2.1. *Data*

Using a single journal minimizes a variety of random forces that might affect the rhetoric of articles. Different journals have different rhetorical conventions. Some prefer short articles, some long. Some demand many references, some prefer few. Some require particular forms of quantitative analyses, others have broader tastes. All of these differences lead to differences of rhetoric—omitting a literature review or expanding a "data" section or whatever.

Of course, rhetorical patterns change over time, even within a single journal. But by selecting a central disciplinary journal, at the top of a discipline's prestige hierarchy, with purported general interests, we can come as close as possible to a situation in which article rhetoric reflects neither journal differentiation nor particular editors, but rather reflects central disciplinary forces. If the journal has enjoyed such a central position over many years, these forces will play on it in a relatively constant manner. Rhetorical changes will not arise in changes of the journal's status *relative* to other journals.

The *American Journal of Sociology* (AJS) meets these criteria perhaps as well as any journal in social science. It has been central to its discipline for 100 years. Its editors—only ten in that period—all came from the same department, whose relative status in the field changed only marginally over the period. It is true that other sociology journals began to appear after World War I and that the *American Sociological Review* cut seriously into AJS's market after the 1930s. But these new journals did not substantially change AJS's character as a central, high-prestige, general-interest journal. It is also true that sociology was hardly a "scientific" discipline in the early years of the journal, before 1925. But the turning of the discipline toward science is precisely one of the things we hope to investigate and thus this fact is less a disability than an opportunity.[3]

AJS thus stands in a roughly consistent position throughout the period in which the "modern sociological article" supposedly emerged (1900–1960). Its articles form a useful universe for our investigation.

[3]One referee pointed out that the AJS often tried to step outside the normal genres of articles, referring us to editor Charles Bidwell's call for unusual papers in July 1976. Having just completed a monographic history of the AJS, however, the senior author can report that it was precisely the feeling that submissions were homogeneous—intellectually as much as or more than rhetorically—that drove Bidwell, as well as his successors, to take drastic measures to find interesting work.

2.2. *Sampling*

Our sampling strategy was driven by the need for adequate representation from each time period and by balanced representation of the different types within the general category of empirical articles. The overall time period was set at 1895–1965. By the latter date, as we noted, the current rhetorical format of empirical articles seems by general account to have stabilized. For example, it is around this time that one first sees complaints about the formulaic character of such articles (Abbott 1996a). At the other end of the period, it seemed best to begin with the first years of the journal, even though empirical reports make up a relatively small fraction of its contents at that time; it is crucial to trace the modern rhetoric to its roots.

The 70-year index of the AJS (*American Journal of Sociology* 1966) codes articles into eight types: case analysis, general quantitative, experimental, survey, cross-cultural, methodological, nonquantitative, and theory. The first four of these can be taken as the relevant universe of scientific/ empirical articles, those whose main burden is to present new empirical material.[4]

In order to balance the four specific types within the larger category of empirical articles, we took a stratified quota sample. We aimed at 15 articles per decade, distributed evenly across the four types. Since the experimental articles began only in the journal's fourth decade, this means five each of quantitative, survey, and case analysis articles until that decade, with about four each for the four types thereafter. A number of articles proved unusable for various reasons (extended formats, two-part articles, and the like), and so our eventual sample was 99 (not $7 * 15 = 105$) articles. (The exact numbers of articles per decade were 10, 15, 12, 15, 15, 15, 17.)

---

[4] Some commenters on this paper have wondered about various of these exclusions—why we left out cross-cultural studies, methodological studies, and nonquantitative studies. Our reasons are straightforward. Nonquantitative studies was a residual category, not a positive one. Methodology studies clearly have a different rhetorical structure than "normal research reports," which was the rhetorical form whose emergence we had hypothesized. (Their aim is to introduce new methods, which takes different rhetorical moves.) Finally, cross-cultural studies was also something of a miscellaneous category. We felt that we would increase our likelihood of identifying particular rhetorical structures by focusing on the most likely kinds of papers, understanding that these made up only half of the journal's articles even at the end of the period. Of course, these were not the *only* forms of articles. No such claim is made here.

2.3. *Coding*

2.3.1. *Reasons for Formal Coding*

Our third major methodological choice was for a formal coding procedure. There are of course the usual important reasons for coding: consistency across cases, disciplining of reflection, and so on. Thus Bazerman's detailed reading of three political science articles (1988:c. 10), however subtle, is vulnerable to the idiosyncrasies of the particular articles chosen and is in many ways more illustrative than definitive. Moreover, since it was a purely cross-sectional analysis devoted to finding "a version of rhetoric," Bazerman's chapter had no need to create an analytic structure that would enable measurement and comparison across time. His equivalent over-time discussion (1988:c. 9) makes length of article its main indicator of rhetoric and contains numerous statements about quantities and proportions of subsections in psychological articles (e.g., 1988:272). But it has no basis in formal measurement, relying rather on the informed judgment of a recollecting reader.[5]

One obvious source for the belief that sociological articles have become more formulaic is change in the proportions of articles using any version of a scientific/empirical format. By the coding used by the AJS indexers in 1965 (*American Journal of Sociology* 1966), the proportion of these articles went from around 15 percent of the total in the first three decades of the journal to 35 percent in the second three decades and to 55 percent by the seventh decade, since which it has steadily increased. Our concern, however, is not with change in the mix of types of articles, although such change undoubtedly reinforces perceptions of change in "the typical article" across the discipline.[6] Rather, we are concerned with change in the rhetorical structure within that group of articles that in particular present some form of systematic empirical information. For it is within such articles that the supposed rhetorical

---

[5]Bazerman's book is meant to open a discussion rather than settle it, and so there was no need—scientific, or, if we wish to follow his discussion, rhetorical—for his interpretations and conclusions to have the support required for one to believe them firmly.

[6]There is good evidence on various changes in topics and areas in sociology journals over time. See Buehler, Hesser, and Weigert (1972), Champion and Morris (1974), Assadi (1987), Kinloch (1988), Murray (1988), and Garnett (1988). See also the various papers in *The American Sociologist*, 11:3, 1976, an entire number devoted to sociological journalism.

straightjacket has emerged.[7] But at a distance, these empirical articles are often indistinguishable. Only by coding them in some detail can we see changes in the nature of their form.[8]

2.3.2. *The Coding Procedure*

Like most coding schemes, ours went through a number of iterations. Our initial decisions concerned the level at which to code units of argument. We knew from the outset that the different genres (theory, quantitative, etc.) have quite different overall structures. We thus followed Barthes (1974) and began at the lowest level of coding, the sentence. We began building up a simple list of types of statements: introduction of a question, statement of significance, empirical assertion, statement of position, identification of elements of an argument, summary of an argument, and so on.

This sentence-level investigation produced a curious discovery in a number of early articles. Papers in the first two decades of the journal sometimes have mixtures of data, theoretical statements, and policy assertions within the same paragraph. This form of writing is relatively more common in the "social policy" articles, but occasionally appears in the kinds of empirical articles analyzed here. This style may well have emerged from the homiletic style of the time, for which it was a standard rhetorical strategy. (It should be recalled that sociology was considered a

---

[7] Of course, in an ultimate sense, these levels of analysis cannot be firmly distinguished. It is possible that the clarification of the rhetorical form of the empirical article helped lead to the domination of journal's contents by such articles.

[8] Coding was also necessary because of our desire to apply formal methods for comparing the various articles. One *could* pick a sample of early and late articles, compare them hermeneutically, and argue that the obvious differences betrayed a tendency toward rigid form. (This was more or less Bazerman's procedure.) But this would not tell us exactly wherein those differences lay. Nor would it, indeed, tell us that there was movement toward rigidity rather than, for example, change from one rigidity to another. Without a formal method for comparing the articles, such questions could not be precisely answered. There are, to be sure, methods for quantitative comparison that do not involve actually coding the individual steps in each article's argument. For example, one could study length in the belief that shorter articles mean more formulaic rhetoric or study the degree of reference to the work of others, defining citation as an emerging rhetorical device. Both these strategies were used by Bazerman. But length is a very imprecise measure of rhetoric. And citation measures are largely dependent on the numbers of other journals available to cite, on the libraries available to citing authors, and other such environmental factors. (Indeed, citation is even dependent on printing technologies; it has ballooned since the rise of the personal computer.)

religious field at the time.[9]) Fortunately, this style was not common in the more empirical pieces. Otherwise, higher level coding would have been meaningless.

We were interested in higher-level rhetorical structures and hence decided on paragraph coding. Changes in lengths of typical paragraphs over the 70 years investigated led us to also code the data in terms of fractions (quarters) of pages for each type of writing. The two systems produce more or less equivalent results, and we report the page coding here, which we regard as more stable over time. Thus the basic unit of coding is the quarter page. Note that this does not mean grid coding—taking each quarter page and asking what it is that appears there. Rather, we first separated an article into distinguishable elements and then measured the length of each element in quarter pages (or paragraphs, in the other metric). Note that both of these metrics preserve information about the lengths, as well as the order, of elements.

For contents, our initial coding began from the same common-sense understanding of social science articles that is discussed in Bazerman (1988 c. 10). The empirical article is generally made up of introductory material, hypotheses, literature, data and methods, analysis, and concluding material. We initially distinguished 28 subtypes under these general article elements. These are shown in Table 1. They are grouped in categories under the overarching headings just given. The elements are defined verbally in the right column. Each element has a basic ID code, listed at the left. We also give the number of times an element was observed in the complete data set ("Number in state" column). By "number of times" we mean here the number of separate episodes or runs, not the number of individual quarter-page units. (Otherwise length of articles would be the dominating force in these counts.) Thus a run of 30 units of one element in one article counts here as one observation of that element.

In coding, it soon became clear that these elements are often combined. Indeed, sometimes several codes are combined. In Table 2, we list the various combination states. At the left are listed the ID codes of the simple states constituting each combination state. We then list the number

TABLE 1
Elementary Elements of Articles

State	Number in state	New ID	
A. Introductory			
11	2	1	Topic
12	66	2	State of affairs
13	6	3	Question
14	6	4	Language
15	13	5	Significance of paper
16	11	6	Purpose of paper
17	54	7	Author's theory/assertion
18	9	8	Existing discussion
B. Hypotheses			
21	29	9	Existing theory/hypotheses
22	22	10	Author's hypotheses
C. Literature			
31	2	11	Citation of existing author
32	8	12	Summary of argument
33	11	13	Commentary on existing author/ argument
D. Data and Methods			
41	14	14	Sources of data
42	71	15	Description of research design
43	28	16	Empirical veracity
44	25	17	Methods of analysis of data set
E. Discussion of Data/Analysis			
51	16	18	Presentation of figures
52	5	19	Report of analysis of data set
53	23	20	Summary of findings
54	39	21	Total summary of findings
55	13	22	Relevance for theory/hypotheses
56	16	23	Total relevance for theory/hypotheses
57	11	24	Identification of cause or explanation
58	14	24	Identification of cause or explanation
F. Conclusion Material			
61	3	57	Review of findings
62	18	58	Implications
63	10	59	Prescriptive statements

of times each combination state was observed in the total data set, with the same conventions about runs as in Table 1.

To simplify final data presentation, we assigned new ID codes to all states, simple and combination. These are shown in the third column of each table. Note that the original codes of the simple states were chosen so that their tens digit described their general category and their units digit a specific type. This makes the combination states a little easier to understand. A few complex combinations were observed only once or twice. In the data here these are not reported. In analysis they were lumped under a residual element, new ID 56.

2.3.3. *Aggregate Trends in Coded Units*

Some comment is necessary on the empirical trends in the elements observed in the 99 articles. First, there were no obvious trends by decade in the total number of counted units per decade, in the relative proportion of combination units, (although there was perhaps a slight tendency to more complex combination units in later years), and in the tendency to use more than one kind of combination unit. Moreover, the median number of units per article is strikingly constant at 10, as are the first and third quartiles of that distribution (at 8 and 13 respectively). The reader should recall that this means *runs* of identical elements, not actual length in numbers of quarter pages; the duration information will be used later in analysis.

The lack of change in combination states argues against one form of rhetorical rigidification: the distillation of article sections into simple, single-purpose units. One possible rhetorical change might have been from the homiletic style (with its melange of facts, analyses, and interpretations) to a style of mixed units and thence to a style marching through rigidly purified sections of theory, methods, data, analysis, interpretation, and conclusion. We do not find evidence consistent with this directional change.

There are minor trends among types of units, which begin to suggest some empirical trends in rhetorical patterns. There is a move from presenting simple numbers and quantities to presenting formal data analysis, signified most clearly by a decline over time in element 40 (presentation of figures plus summary of findings) and a corresponding increase in element 43 (data analysis plus summary of findings). A similar pattern appears for the elements 51 and 53, which are simply elements 40 and 43 plus a "discussion of relevance for theory and hypotheses." Second, there

TABLE 2
Combination Elements of Articles

State	Number in state	NEW ID	
Double States:			
12 17	2	25	State of affairs//Author's theory/assertion
12 51	1	26	State of affairs//Presentation of figures
12 53	2	27	State of affairs//Summary of findings
17 51	4	28	Author's theory/assertion//Presentation of figures
17 53	7	29	Author's theory/assertion//Summary of findings
17 54	10	30	Author's theory/assertion//Total summary of findings
17 55	3	31	Author's theory/assertion//Relevance for theory/hypothesis
17 57	1	32	Author's theory/assertion//Identification of cause or explanation
17 58	1	32	Author's theory/assertion//Identification of cause or explanation
22 32	2	34	Author's hypothesis//Summary of argument
22 55	3	35	Author's hypothesis//Relevance for theory/hypothesis
31 32	15	36	Citation of existing author//Summary of argument
32 33	2	37	Summary of argument//Comment on existing argument
41 51	1	38	Data sources//Presentation of figures
43 51	2	39	Empirical veracity//Presentation of figures

51 53	74	40	Presentation of figures//Summary of findings
51 55	5	41	Presentation of figures//Theory/relevance
51 57	0	42	Presentation of figures//Identification of cause or explanation
52 53	45	43	Data analysis//Summary of findings
53 55	10	44	Summary of findings//Theory/relevance
53 58	2	45	Summary of findings//Identification of cause or explanation
54 55	2	46	Total summary of findings//Theory/relevance
54 56	7	47	Total summary of findings//Total theory/relevance
Triple and Quadruple States			
17 51 53	4	48	Author's theory//Presentation of figures//Summary of findings
21 53 55	2	49	Existing theory//Summary of findings//Theory/relevance
22 52 53	1	50	Author's hypothesis//Data analysis//Summary of findings
51 53 55	14	51	Presentation figures//Summary of findings//Theory/relevance
51 53 57	5	52	Presentation figures//Summary of findings//Identification of cause or explanation
52 53 55	16	53	Data analysis//Summary of findings//Theory/relevance
52 53 58	2	54	Data analysis//Summary of findings//Identification of cause or explanation
52 53 55 58	1	55	Data analysis//Summary of findings//Theory/relevance//Identification of cause or explanation

is a substantial increase in general category D (data and methods) elements, specifically in element 15 (description of research design) and element 17 (methods of analysis of data set). As Bazerman notes with respect to psychology, these sections provide crucial rhetorical anchors in a claim to scientific authority. The decadal totals for all these elements, shown in Table 3, suggest that even these trends are not absolutely regular. Nonetheless they do hint at a move from simply reporting numbers to doing something with them inferentially and date that move pretty clearly to the fourth and fifth decades of the journal, the period 1925–1945.

3. OPTIMAL ALIGNMENT OF FULL ARTICLES

We now turn to an analysis of the exact rhetorical patterns within articles. At this point, we reintroduce information on the exact duration of particular sections of articles, which we had removed during the aggregate analysis in order to somewhat standardize for the effects of changes in article lengths. (That standardization is done differently in the alignment analysis below.)

A typical data sequence is the following:

1A2   1A9   2A15   20A43   5A21   5A23

The letter A means nothing, but merely separates the number of times an element is observed sequentially, which precedes the letter, from the identity of the element, which follows it. This sequence thus means that the article comprises, in the given order:

  1 unit of "state of affairs" followed by
  1 unit of "existing theory and hypotheses" followed by
  2 units of "description of research design" followed by
20 units of data analysis/summary of findings followed by
  5 units of total summary of findings
  5 units of total (overall) discussion of relevance for theory
     and hypotheses

3.1. *Optimal Alignment*

Given such codings for the 99 articles in the sample, we sought sequence patterns by using optimal alignment methods, which have been described in detail elsewhere (Abbott and Hrycak 1990; for a general reference see Sankoff and Kruskal 1983). Alignment methods are simple replacement

TABLE 3
Decadal Observations of Trending Elements

Element	Decade						
	1 1895–1905	2 1905–1915	3 1915–1925	4 1925–1935	5 1935–1945	6 1945–1955	7 1955–1965
15 Description of research design	5	3	8	12	14	13	16
17 Methods of analysis of dataset	1	1	2	5	7	6	3
40 Presentation of figures//Summary of findings	11	27	12	18	4	1	1
43 Data analysis//Summary of findings			1	1	16	18	9
51 Presentation of figures//Summary of findings// Theoretical relevance		4	3	6	1		
53 Data analysis//Summary of findings// Theoretical relevance				1	6	3	6

algebras that compute a distance between any pair of sequences based on the minimum number of replacements and insertions required to transform one of the sequences into the other. For example, it takes two changes to turn the sequence 1A2   1A9   2A15   20A43   5A21   5A23 into the sequence 2A2   1A9   1A10   1A15   20A43   5A21   5A23. These changes are easily seen by aligning the sequences, that is, arraying them so that their similarities are lined up:

1A2	1A9		2A15	20A43	5A21	5A23
2A2	1A9	1A10	1A15	20A43	5A21	5A23

A unit of A2 has been added, and a unit of A15 has been turned into a unit of A10. One can easily imagine a simple metric for assessing distance between pairs of sequences, consisting of the number of such insertions and replacements required for the full transformation. Note that there are other ways to make this change, but they involve more total insertions and replacements. One could insert the A2, then insert an A10, then remove an A15, for example. This would allow alignment with only one elementary operation (insertion/deletion), where we have allowed two (insertion/deletion plus replacement). But it takes three insertions/deletions where we have used one insertion/deletion and one replacement. (In some circumstances, however, there might be reasons for not allowing replacement as an operation.)

There is a variety of alignment algorithms, depending on what particular kinds of sequence regularities are of interest.[10] The algorithms vary, first, in the elementary operations used (insertions, replacement, swaps, transpositions, and the like) and, second, in the way costs are assigned to those operations. The example just given implicitly assumes that insertion/deletion "costs" the same thing as replacement and that all replacements "cost" the same. Under some circumstances those assumptions might not be useful. Algorithms may also choose to regard any string of insertions—however long—as having a given, fixed cost.

In the present case, we are agnostic about the kinds of regularities we expect. The regularities could be local patterns—subsections that al-

---

[10] The senior author has briefly noted some of these different algorithms in a comment on a paper in *Sociological Methods and Research* that introduced a variant algorithm as if it were a completely new method (Abbott 1995b). A review of the state of the art in biology (as of 1991) was Gribskov and Devereux (1991) and a more recent (slightly) review is Boguski (1992). Developments in this area, which is force-fed with biotechnology money, are very rapid.

ways come in the same order. They could be "stage" regularities, where certain rhetorical elements are generally early, middle, or late in the article, but the detailed ordering of elements can vary. Where we are so unsure of expected regularities, it makes sense to use a relatively simple algorithm, one with only the two operations of insertion/deletion and replacement and one having fairly simple cost assumptions.

The standard such algorithm is the Needleman-Wunsch (1970) algorithm. It uses the two elementary operations of replacement and insertion. For each, it requires that we assign a "cost" scheme, on the assumption that some replacements and insertion/deletions matter more than others.

Replacement costs have to be set for all pairs among the 59 potential sequence elements listed in Tables 1 and 2. For the "simple" elements (those in Table 1), replacement costs were set at 1.0 for all elements falling under different headings in that table, and at 0.25 for all such units within a given heading. Thus substituting element 15 for element 16 cost only 0.25, but substituting element 15 for element 20 cost 1.0. This ratio (0.25 to 1.0) is somewhat arbitrary; it could easily have been 0.33 to 1.0, for example. But it catches our sense that emphasizing different whole sections of an article is a more drastic rhetorical change than changing the emphasis among ways of writing a particular section.

For the combination elements (elements 25–55, in Table 2), costs were set at the average of replacement costs over all potential pairs of individual elements within the combinations. Thus substituting element 43 (combination of simple elements 52 and 53) for element 26 (combination of simple elements 12 and 51) cost an average of 1.0, 1.0, 0.25, and 0.25, or 0.625. This weighting again made the algorithm tend to emphasize closeness within the element heading types where perfect matches were not possible.

For insertion costs, we chose a uniform but very low value. Many biological alignment algorithms use a fixed "gap penalty" for any insertion plus a much smaller weight linear in the length of the insertion. We chose a uniform insertion cost—that is, an insertion cost scheme with no gap penalty and a fixed insertion/deletion cost linear in the number of inserted elements. We choose this because we feel that the insertion involved in article rhetorics is more a matter of gradual expansion or contraction of subsections than of insertion of new blocks of random length. At the same time we have set the insertion cost quite low. Insertion costs that are high—more than half the largest or even the mean replacement cost—tend to force the alignment to "follow the main diagonal"— that is,

to use replacements preferentially, because it takes two insertion/deletions to make one replacement. This in turn makes differences in length unduly important in the alignment, even when, as here, the intersequence distances are all divided by the length of the longer sequence in order to standardize for such differences. An insertion/deletion cost that is low relative to replacement cost makes the algorithm put more emphasis on the order pattern of the elements than on the duration of runs of elements. We have set insertion costs at 0.1 times the largest replacement cost, which means that where replacement costs are high, the algorithm will handle replacements as combinations of insertions. In practice, this means that the algorithm is free to focus on aligning perfectly those parts of the sequences that very closely resemble one another, using insertions more or less as necessary to find these "common locations."[11]

3.2. *Analysis and Results*

Using these parameters, the data were input to an optimal alignment algorithm. The algorithm produces a distance between every pair of articles in the data—99 ∗ (98/2), or 4851 pairs. For a first judgment of the hypothesis

---

[11] All calculations in this paper were done using the program OPTIMIZE. The program (for DOS, Windows, and OS2) is available on the Internet at Web site http://www.cicero.spc.uchicago.edu/usurs/abbot. OPTIMIZE is a simple program for doing optimal alignments following the Needleman-Wunsch algorithm as adapted here. It has two modules, one of which (EXPLORE) permits visual inspection of the actual algorithm pattern between any pair of sequences and allows the investigator to rapidly scan the impact of varying the insertion costs on the actual alignment observed. There are literally dozens of other alignment programs available, most designed for biological applications and input (see Gribskov and Devereux [1991]).

There is at present a small literature on the impact of varying insertion costs, usually in the context of a single-cost replacement model. If one defines a Cartesian space by insertion cost along one axis and replacement cost on another, different points in the space may produce different alignments, in the literal sense of a different sequence of elementary operations (or the same operations at different places), between two given sequences. These changes induce a tessellation of the space, one that becomes even more complicated when we introduce varying replacement costs and with them more dimensions of the space. (See Vingron and Waterman 1994.)

Given the preliminary nature of the formal literature, we must fall back on the test of practice. Forrest and Abbott (1990) present an extended reliability analysis of optimal alignment of coded data, showing that fairly large local differences in coding procedures still produce recognizably similar results because the gross data are constant. In working with various data sets, the senior author has found that placing the insertion cost at 0.1 times the largest replacement cost is usually a good starting point. The aim of alignment is to find patterns of interest. This level seems in practice to produce the best results. It is important, however, to test a variety of values.

of convergence, we calculated means and standard deviations for these distances within and between decades. These are reported in Table 4. (Thus, 0.682 is the mean distance between all pairs of articles such that both were published in the first decade. 0.159 is the standard deviation of this figure.) It is clear by inspection that the data fall into two broad groups: decades 1–4 on the one hand and decades 6 and 7 on the other. Decade 5 is clearly transitional but leans toward the later group. This division into two groups is largely due to the shift from figures to data analysis noted in Table 2. Regardless of this trend, there is no evidence whatever in this table that intradecade distances decrease with time, as would be the case if convergence were occurring.

TABLE 4
Distances Within and Between Decades

A. Mean Distances

	Decade						
	1	2	3	4	5	6	7
	1895–	1905–	1915–	1925–	1935–	1945–	1955–
Decade	1905	1915	1925	1935	1945	1955	1965
1 1895–1905	.682						
2 1905–1915	.641	.590					
3 1915–1925	.633	.585	.613				
4 1925–1935	.640	.607	.639	.621			
5 1935–1945	.716	.724	.744	.719	.735		
6 1945–1955	.753	.774	.776	.749	.707	.630	
7 1955–1965	.709	.727	.746	.731	.700	.653	.655

B. Standard Deviations

	Decade						
	1	2	3	4	5	6	7
	1895–	1905–	1915–	1925–	1935–	1945–	1955–
Decade	1905	1915	1925	1935	1945	1955	1965
1 1895–1905	.159						
2 1905–1915	.173	.165					
3 1915–1925	.185	.184	.220				
4 1925–1935	.140	.146	.161	.138			
5 1935–1945	.125	.148	.136	.152	.148		
6 1945–1955	.099	.108	.100	.120	.161	.196	
7 1955–1965	.107	.117	.106	.129	.148	.161	.166

These general figures, however, could misrepresent matters if there were multiple forms within particular periods. The earlier period might have been characterized by one form around which there were many minor variants, while the later period might have had two or more forms with much less variation around them. This could produce the distance pattern shown, even though there *had* been convergence. To rule out this possibility, we must consider the data in terms of rhetorical forms themselves.

To search for rhetorical forms, we analyzed the distance data with standard cluster analysis algorithms. In order to avoid the common volatility of cluster analysis results with different algorithms, the sequence distances were analyzed using both single and average linkage. These produced strongly similar results (a somewhat unusual event in our experience). This similarity indicates that the results are fairly robust.[12]

Data for two average-linkage clusters are shown in Table 5. Note that left-right distance has little meaning in this table. We have spread the sequences out to show their alignments most effectively within cluster. (This probably represents that actual alignments used, since insertion/ deletion cost was low.) It is clear how the alignment works—picking up

---

[12]Cluster analysis is a general method for producing categorizations of data based on distance metrics. For a general text, see Everitt (1993). As is well known, different algorithms for cluster analysis produce slightly different solutions. We chose both single linkage and average linkage—the first because despite its tendency to stringy clusters it has the advantage of producing the same clustering under any monotonic transformation of the data; the second for its more stable, globular clusters. As noted in the text, stability of clusters across algorithms is good evidence of substantive validity. We report 14 clusters (rather than, say, 3 or 30) because that level of clustering seemed mostly clearly interpretable.

We have not undertaken direct validity measures of these partitions. On the one hand, there were no bases for external validity criteria (no second and uncorrelated types of measurement of resemblance between articles). On the other, internal validity measurement (validity analysis based purely on the proximity matrix itself) is, in the words of Jain and Dubes, more or less a "black art accessible only to those true believers who have experience and great courage" (1988:222). The analysis here produces credible substantive results, showing a clear and plausible relation between rhetorical structure and period of writing. The clusters persist across different clustering algorithms. Any further validity analysis would take us into obscurities that would prevent our getting on to the matter of subsequence analysis.

There is a recent literature on weighting in sequence space prior to analysis. This mainly aims at the problems of redundancy (Henikoff and Henikoff 1994) and direct interdependence (Vingron and Sibbald 1993) among sequences. The second of these is impossible in this data; the structure of one article does not directly coerce another the way an ancestor DNA can coerce its descendant. The first is important, but since our main purpose is to find the major types of articles, weighting in favor of the "rare article type" seems mistaken. For a general review of statistical issues related to sequence analysis, see Karlin et al. (1991).

TABLE 5
Two Clusters of Sequences

Cluster 1

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
86			1A2	1A8	2A5	2A7	9A19	9A40	9A19	9A40	2A22	18A40	5A21		5A59						
67		3A8	3A40	2A25				17A40	1A20	7A40	6A51		3A21	3A7	2A17						
24	1A1	1A7	1A2			1A17		19A40	5A24	4A40											
52			1A2	1A3		2A15		17A40	2A18	5A41	9A40	2A24	2A21								
15	2A15	1A4	9A40	2A7				13A40						1A7							
37	5A9		1A2			5A15		25A40	1A30												
62			1A2			5A15		27A40	1A18	6A40	5A18										
36			2A2	2A17		2A15		28A40	1A58												
64			1A2	1A9		2A15		12A40	2A24	24A40			2A21	2A23							
88			1A2	1A14	1A16	13A15		49A40	9A16												
80				2A7		6A15		49A40						2A7	2A59						
82			2A2	6A28		2A14		37A40	2A7	13A40				1A7							
81	6A9					2A15	1A17	39A40	6A18	7A40	2A5										
90			9A2	9A18	1A5	5A20		33A40	1A7	18A40				2A7							
95	1A8	1A9	5A2					18A40	2A24	48A40	13A46										
98			2A2			6A20		25A40	6A48	23A40	17A48	10A40	2A21	2A7	2A24	2A59					
97			1A2	1A4		6A18		13A40	5A29	26A40			9A24	30A48	2A56						

Cluster 13

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
53	1A6	1A2	9A15		9A16	5A43	2A17			1A43	1A56	2A54	5A43				5A21				
16	1A6	1A5	2A15	2A12		5A43	2A16			6A43	1A19		1A21	1A9	2A43	2A31	1A53	1A21	1A16		
40	3A56	3A9	1A15							19A43				1A23	6A16	2A56	1A5				1A58
38	2A12	2A6	2A10	2A4	6A15					16A43			2A21	2A23							1A58
41	1A12	1A13	2A2		6A15					17A43			3A21	3A23	3A7		2A57				
69	1A2		14A15			5A43	5A56			15A43			5A21		2A7	1A10					1A58
32	1A2	2A7	10A15		1A17	14A43				5A31											
17	1A2	1A16	1A36		6A15	1A17	1A16			16A43				2A30							
45	1A2		5A15	1A17	1A20	23A43				1A30				5A16	1A5						
49	1A6	1A2	1A9	2A10	2A15	2A39				22A43				6A23							
51	1A2	1A34	5A15							31A43			2A21	2A23							
42	2A2		1A15							32A43			2A21								
70	1A2		1A15			21A43	2A21	2A2	2A15	17A43			5A21								
73	6A2		1A15	1A3		21A43	2A7			20A43					2A7						
79	1A2	5A17	2A15			10A43			1A7	22A43					5A7	11A43	1A30				1A58
85	1A5		3A15			15A43			5A56	39A43			2A21		1A7						
31	5A43	5A27				12A43	2A24			5A43					1A7	6A2					
54	1A2	1A14				2A43	9A24			6A43	6A29	14A43	2A21	1A24							
25	1A2	1A9	2A15	5A53	5A17	11A43	5A17			2A43	1A5	1A14									1A58

various parts of various sequences and lining them up. There is a fair amount of minor variation, but in each case the algorithm has found a basic pattern.

Table 6 presents descriptive statistics on the clusters that clearly persist in both clustering methods. For the major clusters, we report their "mean volume year"—that is, the average of the volumes in which the cluster members appeared. We also report the year represented by this volume (always 5 less than the mean volume year because the AJS began publication in 1895) and the standard deviation of that mean.

The principal result is straightforward. The clusters listed in Table 6 are more or less of equal cluster diameter. Most of them are quite clear and distinctive groups. As we might expect, it is at once clear from the standard deviations that article patterns are not randomly distributed through the 70-year period, but rather that certain article patterns are very clearly as-

TABLE 6
Major Sequence Clusters

Cluster Number	Number of Sequences in Cluster[a]	Mean Volume[b] (Year)	SD[c]	Exemplary Sequence[d]
1	17	20 (1915)	9	9A2 9A18 1A5 5A20 33A40 1A7 18A40 2A7
2	9	30 (1925)	12	2A36 2A15 2A16 16A40 2A21 2A57 2A23
5	6	22 (1917)	11	1A2 2A7 1A6 22A40 4A9 2A10 36A51 21A59
7	5	26 (1921)	27	3A2 3A7 2A14 2A16 20A52 2A40 5A21 9A24
9	9	42 (1937)	21	1A13 1A7 5A36 9A56 7A53 1A21
11	6	44 (1939)	22	1A2 1A3 2A9 2A10 2A17 3A53 1A2 1A36 2A7 3A43
12	9	54 (1949)	11	9A36 6A10 2A15 19A53 6A47 2A58
13	19	55 (1950)	9	1A12 1A13 2A2 6A15 17A43 3A21 3A23 3A7 2A57
14	3	64 (1959)	3	6A12 2A10 2A12 5A15 6A43 2A21 2A23

[a]The number of sequences located in the cluster.
[b]The mean of the volume numbers of the volumes in which cluster members appeared.
[c]Standard deviation of the mean volume figure.
[d]A sequence illustrating the characteristic pattern, so far as there is one, of the cluster.

sociated with particular periods. In particular, there appears to be a vocabulary of forms that were used early (clusters 1, 2, 5, and 7), a vocabulary of forms that were used late (clusters 12, 13, and 14), and some patterns that cross much wider periods (clusters 9 and 11). These data, too, are not compatible with the standard hypothesis of increasing concentration on a particular form. Rather these data suggest a move from one set of standard forms to another.

We first discuss clusters 1, 2, 5, and 7, the dominant early forms. Cluster 1 is clearest. It begins with a stage setting description of a state of affairs (element 2) and sometimes has a prefatory section on research design or analysis (element 15 or 17). It continues through a large section mixing numbers and findings (element 40), an interruption or temporary conclusion (variety of elements), and more figures and findings (element 40) to come to a conclusion, either a simple summary (element 21) or the author's theory (element 7). The cluster 2 pattern is much the same, but it lacks the repeated analysis section, more uniformly possesses the research design section, and sometimes has a section before the main analysis that mixes discussion of theory relevance in with the figures and findings (element 51). Cluster 5 moves further in this direction. In it the introductory sections have come apart, and the mixed section (element 51) now dominates the article, although it is sometimes preceded, not followed, by an analysis section with figures and findings but no intermingled discussion of theory relevance (element 40). Cluster 7 has a clear state of affairs introduction (element 2), usually has other introductory material (discussion of author's or others' theories), then moves straight into a short analytic (element 40) midsection, and often concludes with some simple material summarizing findings (element 21).

For the late period, the main forms are shown in clusters 12, 13, and 14. These are distinctive and clear groups. Cluster 12 begins, in six of nine cases, with a literature review—freestanding (element 13) or combined with a summary of the new argument (element 36). There follows a description or methods section (elements 15–17), and a substantial mixed analysis section (element 53). This is followed by a conclusion that is either a restatement of the author's theory (element 7) or a discussion of implications (element 58). This is close to the formulaic pattern described by Bazerman (1988). Note that this cluster marks the first serious appearance of literature reviews as part of consistent style.

Cluster 13 begins with a state of affairs discussion (element 2), followed in most cases by a data/methods setup section (elements 14–17). Most often this is element 15 (description of research design), but to this is

often added a specific description of methods (element 17). This material is then followed by the main analytic section (element 43), which is in turn followed by some combination of summary of findings (element 21), relevance of findings for theory and hypotheses (element 23), and author's hypothesis (element 7). Implications discussions (element 58) occur occasionally. This version of the contemporary article is more elaborate than the cluster 12 version, but emphatically lacks the earlier version's focus on the literature.

Finally, cluster 14 with its three late but extremely similar articles, shows a slightly different format. Here, we have a summary of the arguments in the literature (element 12), a research design (element 15), in two cases a discussion of the solidity of the data (element 16), a relatively short mixed analysis section (element 43), in two cases a summary of findings (element 21) and a discussion of relevance of the findings for theory and hypotheses (element 23). This is indeed a formulaic pattern; it is quite extraordinary to find three randomly chosen articles whose patterns match so well.

The two clusters of intermediate date both have large standard deviations, which indicates that their intermediacy does not indicate a "transition form" used in a move from earlier forms to later ones, but rather that they are rhetorical structures that appeared throughout the period studied. The first of these is characterized by a large section of element 56, the residual element, which appears only in scattered places outside this group. Since element 56 can stand for a variety of complex combination states (all of which occur only one or two times apiece), there is no guessing what it stands for. Inspection reveals, however, that group 9 does not in fact have a very clear rhetorical pattern. Cluster 11 does, but it is an odd one. The cluster includes five articles, four of which are characterized by short, choppy sections in no particular order. This is clearly a group of rhetorical failures.

4. RHETORICAL SUBSEQUENCES

Another way of looking at rhetorical change focuses on shorter sequences. Perhaps certain parts of articles always follow in the same order. In particular, it might be helpful to focus on order exclusively, disregarding duration. We can then ask whether rhetorical rigor emerges in the form of a particular sequential armature around which an article must be constructed.

Disregarding duration is not difficult; we simply rewrite the data without allowing any repeats; a run of an element, however long, is rep-

resented by one element. But the common subsequence problem itself is more difficult. The technical problem is to find the common subsequences in a sequence data set without knowing ahead of time what those sub-sequences are, where they start, or how long they are likely to be. Fortu-nately, a recent iterative algorithm developed by Lawrence et al. (1993; hereafter, LEA) solves this problem.

4.1. *Gibbs Sampling*

The LEA algorithm employs a Gibbs sampler. Before discussing the algo-rithm, it is important to clarify what Gibbs sampling is and how it works. Tanner (1991) distinguishes statistical methods into "observed data meth-ods," which apply directly to distributions of observed data, and "data augmentation methods," which seek those distributions via more round-about means. In this view, Gibbs sampling is a data augmentation tech-nique, one that constructs existing probability information into an iterative sampling structure that gives access to distributions that are themselves inaccessible via direct means, either because of analytic intractability or computational inaccessibility.

Put most simply, Gibbs sampling uses conditional information to produce marginal information. For a simple illustration, we apply Gibbs sampling to a two-by-two table. Suppose we have two two-valued vari-ables and the conditional information relating them but do not have the marginal distributions. (We can solve for these directly, of course, but use the Gibbs sampler because the simple example makes it easy to understand.) Thus we have (1) a row-conditional matrix A showing $\Pr(y|x)$

	$y = 1$	$y = 2$
$x = 1$	.424	.576
$x = 2$	.556	.444

and (2) a column-conditional matrix B showing $\Pr(x|y)$

	$y = 1$	$y = 2$
$x = 1$	.583	.704
$x = 2$	.417	.296

If we start with an arbitrary value for x, we can draw a value for y following the conditionals in the first matrix. We can then use *that* y to draw a value for x following the second matrix. We could then use our new x to draw a value for y following the first matrix and so on. This process will converge, in this case quite quickly, to a random sample on the true x marginals. To see why, consider the following argument.

If we transpose the B matrix, we have a row-conditional matrix with *y*-defined rows and *x*-defined columns:

	x = 1	*x* = 2
y = 1	.583	.417
y = 2	.704	.296

Consider the product A times B-transpose: This is a row-stochastic matrix (that is, its rows each sum to one. By elementary algebra the product of any two conformable row-stochastic matrices is also row-stochastic.) Note that this product can be thought of as giving the probability of *x* given an arbitrary starting *x* used as a basis to choose a *y* under the conditionals given in B-transpose and then a new *x* following those given in A. That is, it gives Pr(*x*) given an *initial x* and a tour through Y:

	x = 1	*x* = 2
x = 1	.653	.347
x = 2	.637	.363

The square of this matrix is again row-stochastic and describes Pr(*x*) given an arbitrary start in X and *two* tours through Y:

	x = 1	*x* = 2
x = 1	.647	.353
x = 2	.647	.353

But both rows are the same, which indicates an equilibrium X distribution: in fact, that of the X marginals. The equivalent argument can be made for the Y marginals. This time, we start with B-transpose then move through A. The product B-transpose times A is:

	y = 1	*y* = 2
y = 1	.479	.521
y = 2	.463	.537

And the square of this matrix is:

	y = 1	*y* = 2
y = 1	.471	.529
y = 2	.470	.530

In this case, too, the matrix ends up quickly with identical rows. Reassuringly, the original joint matrix is in fact:

	y = 1	*y* = 2
x = 1	.275	.373
x = 2	.196	.157

which gives exactly these marginals. An equivalent design could be set up for a three-way table for which we possessed the three two-by-two conditional tables specifying the dependence of each variable on particular values of the other two. Indeed, this algorithm applies to larger two-way tables and indeed to any *n*-way table for which we possess the complete n-1 way conditionals. These conditionals can be used for a direct algebraic calculation of the joint probabilities (and hence the marginals), of course, but the case offers a nice example of Gibbs sampling.

In formal terms, Gibbs sampling defines a transition function on the space of all possible joint values of the variables. Note that all possible joint values could in principle be reached by this function and also that there is no reason to expect the function to get trapped in a particular "cycle" of successive joint values. On these assumptions (irreducibility and aperiodicty) the transition function defines an ergodic Markov chain (see, e.g., Cox and Miller 1965). Since a regular Markov chain has a limiting distribution, it makes no difference what arbitrary values of the variables we start with; the process eventually finds the equilibrium distribution of the variable of interest. Given a system with *n* variables and sets of complete conditional probabilities describing the probability of particular values of each variable given any possible combination of values of all the others, the basic Gibbs algorithm is as follows:

1. Set all variables at arbitrary initial values.
2. Select a new value for variable 1, given the probabilities of that variable conditional on the current values of all the others, which were fixed in step 1.
3. Accept that value for variable 1 and select a new value for variable 2, given the probabilities of variable 2 conditional on the current fixed values of variables 1 and 3 through *n*.
4. Repeat step 3 for all variables and then cycle through the variables again and again until convergence.

In cases less simple than that of *n*-way tables, we often have far less ability to move algebraically from conditional probability to joint or marginal probability. It is in these cases that Gibbs sampling is more necessary. While Gibbs samplers are most commonly used (as here) to develop marginals for intractable distributions, they can also be used for optimization, as in the application given below. Gibbs samplers are part of a larger universe of methods now usually referred to as Markov chain Monte Carlo (MCMC) methods. The oldest of these is the famous Me-

tropolis algorithm for estimating state equations on spaces of individual particles.[13]

4.2. *The LEA Algorithm*

The LEA algorithm uses Gibbs sampling to attack the common subsequence problem. In order to illustrate it, we apply it first to some experimental data with a particularly strong subsequence regularity. The data are twenty random sequences of 50 integers between 1 and 15. At some random point in that sequence, we have overwritten the random digits with the subsequence 8 9 10 X X 13 14 (where X stands for some random number.)

An example is:

11 11 15 9 5 5 12 1 12 13 13 9 7 8 4 10 8 11 14 6 5 5 3 8 9

10 15 14 13 14 11 12 10 12 3 13 4 15 1 1 12 6 7 2 2 3 1 11 8.

The regularity starts at the end of the first line. The data consist of 20 such sequences with their written-in regularities.

The LEA algorithm begins by randomly generating a list of starting points for the common subsection of each data sequence, a procedure that fills a matrix whose rows are the candidates for the common site in each sequence. We shall call this the candidate matrix. (Following the LEA notation, we call this matrix A.[14]). At the start, it is filled with randomly

[13] The Metropolis algorithm was set forth in Metropolis and Ulam (1949) and Metropolis et al. (1953). The classic paper redeveloping the algorithm as an optimization technique is Kirkpatrick et al. (1983). Because of the motivation of these methods by analogy with the physical cooling of metals, the technique of combining probabilistic transition with probabilistic acceptance of that transition even if it worsens an objective function is called simulated annealing.

The Gibbs sampler, like MCMC methods more generally, is part of a growth industry in the statistical community today. A straightforward introduction is Casella and George (1992). A more general and analytic exposition is Liu (1991:c. 6), which also makes detailed comparisons with the Metropolis algorithm. An account combining several classic papers with lengthy comments by dozens of statisticians can be found in the *Journal of the Royal Statistical Society*, Series B, (1993), 55:3-102. The name "Gibbs sampling" was coined, and the new generations of such probabilistic optimization begun, by Geman and Geman (1984). The approach was generalized in statistics by the work of Gelfand (Gelfand and Smith [1990], Gelfand et al. [1990]). A related technique of data augmentation is that of Tanner and Wong (1987). A source locating a variety of data augmentation techniques relative to one another is Tanner (1991). A current review of MCMC methods is Gilks et al. (1996).

[14] Our exposition directly follows that of Lawrence et al. (1993). Simulated annealing has also been applied to the global multiple sequence alignment (GMSA) problem (as opposed to its use here for the multiple subsequence alignment problem) by Ishikawa et al. (1993). Other approaches to the GMSA problem are Gotoh (1993) and Gusfield (1993).

chosen subsequences. An example follows. Each row contains the algorithm's first random guess at the regularity in a particular data sequence. We have guessed that the regularity is seven units wide. (In practice, one would not know this and would have to try several widths to find the best result.) Note that the start is not auspicious; pieces of the "real" regularity appear in only three candidates, those for sequences 11, 15, and 19.

The Original A Matrix
Subsequence

	1	2	3	4	5	6	7
1	9	5	5	12	1	12	13
2	8	8	7	6	7	5	1
3	2	3	11	7	6	3	11
4	12	5	11	4	2	1	5
5	12	6	7	2	2	3	1
S 6	9	8	15	4	6	6	5
e 7	8	3	4	5	12	1	8
q 8	11	7	13	12	7	2	7
u 9	1	13	12	12	5	7	9
e 10	15	10	8	4	9	12	7
n 11	1	3	7	9	8	9	10
c 12	1	15	2	5	9	8	9
e 13	14	6	2	7	12	6	7
14	9	1	5	3	5	12	10
15	6	4	8	9	10	7	7
16	10	1	9	13	14	14	8
17	9	11	5	9	9	15	14
18	1	11	13	8	13	15	8
19	9	10	12	14	13	14	11
20	1	1	6	7	6	10	3

We also calculate the background probability of each element over the entire data set. We call this vector P.

The General P Vector
Elements

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Pr = .06	.05	.06	.06	.08	.06	.08	.08	.08	.08	.06	.07	.07	.07	.06

Each data sequence of 50 elements will in turn be considered the "focus" sequence. For each focus sequence, we go through two phases of cal-

culation. First, we have a "Q-calculation phase." Ruling out for the moment the focus sequence's row in the candidate matrix, we calculate the probability of each element at each point in the common subsequence. That is, we calculate the probability of each element in any given column of the candidate matrix. These are "location specific probabilities:" the probability of element 1 in the first position, the probability of element 1 in the second position, etc. This will be called matrix Q. The (i, j)-th element of Q is the probability that element j is in the ith position in the subsequence. To deal with the problem that elements can disappear completely from a column of the candidate matrix A, this matrix adds in a Bayesian prior for each element at each position. Thus, while there are no occurrences of elements 3,4,5,7, and 13 in the first column of A, the prior still produces a nonzero probability. This probability is "lost" from the elements that *do* occur. Hence the Q(1,1) element is not $5/19 = 0.26$, as would appear from the five occurrences of element 1 in the first position in the last 19 rows of the A matrix above, but 0.22. (For details, see Lawrence et al. 1993:209 and n. 20).

The Original Q Matrix
Elements

Subsequence Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	.22	.05	.01	.01	.01	.05	.01	.10	.23	.06	.05	.10	.01	.06	.05
2	.14	.01	.14	.05	.10	.10	.06	.10	.02	.10	.10	.01	.06	.01	.05
3	.01	.10	.01	.05	.14	.05	.14	.10	.06	.02	.10	.10	.10	.01	.05
4	.01	.05	.05	.14	.10	.05	.14	.06	.14	.02	.01	.14	.06	.06	.01
5	.05	.10	.01	.01	.10	.14	.10	.06	.14	.06	.01	.10	.10	.06	.01
6	.10	.05	.10	.01	.06	.10	.10	.06	.06	.06	.01	.14	.01	.10	.10
7	.10	.01	.05	.01	.10	.01	.19	.14	.10	.10	.10	.01	.06	.06	.01

Next comes what can be called the "sampling phase." Consider the 50-element focus sequence. If the common subsequence is 7 elements long, there are 44 possible choices for candidate site in this sequence. (In general, there are l [= length of sequence] minus w [= width of subsequence] plus one possible choices.) For each of these 44 potential candidates, we calculate its probability given the location-specific element probabilities just derived, and divide this by its probability given the background probabilities. Thus there are 44 possible seven-element subsequences in the first sequence. For each of them, we calculate these two probabilities, simply looking up the relevant values in P and Q and multiplying them out. For each possible subsequence in the focus sequence, the quotient of its probability given Q by its probability given P tells whether that candidate sub-

sequence is "better" in terms of the current list of candidates across all *other* sequences. It is a measure of quality of fit *relative* to the current list.

Now one simply adds these quotients for all the candidate sites across the focus sequence and sets the sum to one, normalizing the quotients and turning them into probability measures. Then one uses *these* probabilities to pick one sample from the candidate sites in the focus sequence and enters that site in the A matrix as the new candidate from the focus sequence.[15]

The iteration consists of repeating the Q-calculation phase and the sample phase for each sequence in turn.[16] After one iteration (Q and sample phases repeated once for each of the twenty sequences), here is the new A matrix.

		Subsequence						
		1	2	3	4	5	6	7
	1	10	10	4	5	8	9	10
	2	14	4	12	6	5	14	10
	3	9	10	9	13	2	3	11
	4	1	8	4	13	9	12	14
	5	11	11	15	6	8	7	11
S	6	9	10	4	8	13	14	11
e	7	1	8	12	13	5	15	8
q	8	15	10	9	14	8	9	10
u	9	4	11	8	3	8	9	10
e	10	10	8	4	9	12	7	14
n	11	9	10	12	5	13	14	8
c	12	9	10	4	5	13	14	12
e	13	1	10	13	10	7	3	8
	14	9	1	5	3	5	12	10
	15	4	8	9	10	7	7	13
	16	9	10	1	9	13	14	14
	17	9	11	5	9	9	15	14
	18	1	11	13	8	13	15	8
	19	9	10	12	14	13	14	11
	20	10	10	15	7	8	9	10

[15] The Gibbs sampler as implemented by LEA differs from simulated annealing (SA) methods. Under an implementation with SA, we would accept the newly "sampled" candidate with probability one if its quality measure was better than that of the current candidate from this focus sequence *and* with some nonzero probability (set by the annealing schedule) even if it were worse. In the typical SA analysis, this probability (analogous with temperature) is set high in the early iterations to avoid trapping in local optima; one accepts a lot of random "worsening." As more iterations are run, the probability falls; one is more certain to have found the global optimum.

[16] We could also do this with random choice (so-called random scan). In the implementation used here, the iteration was for each sequence through the data set (systematic scan). What the methods require is that the scan guarantee that the whole space be visited in finite (although potentially very large) time.

At the third pass,

		Subsequence						
		1	2	3	4	5	6	7
	1	10	10	4	5	8	9	10
	2	9	10	9	1	13	14	4
	3	9	10	9	13	2	3	11
	4	1	8	4	13	9	12	14
	5	9	10	11	14	13	14	11
S	6	9	10	4	8	13	14	11
e	7	1	8	12	13	5	15	8
q	8	15	10	9	14	8	9	10
u	9	4	11	8	3	8	9	10
e	10	9	5	5	13	8	9	10
n	11	9	10	12	5	13	14	8
c	12	9	10	4	5	13	14	12
e	13	1	10	13	10	7	3	8
	14	9	1	5	3	5	12	10
	15	9	10	7	7	13	14	1
	16	9	10	1	9	13	14	14
	17	9	11	5	9	9	15	14
	18	9	10	11	11	13	14	8
	19	9	10	12	14	13	14	11
	20	10	10	15	7	8	9	10

And the fourth,

		Subsequence						
		1	2	3	4	5	6	7
	1	8	9	10	13	8	13	14
	2	8	9	10	9	1	13	14
	3	1	9	10	9	13	2	3
	4	6	1	8	4	13	9	12
	5	8	9	10	11	14	13	14
S	6	8	9	10	4	8	13	14
e	7	12	1	8	12	13	5	15
q	8	2	15	10	9	14	8	9
u	9	8	9	10	3	11	13	14
e	10	6	9	5	5	13	8	9
n	11	8	9	10	12	5	13	14
c	12	8	9	10	4	5	13	14
e	13	8	9	10	7	13	13	14
	14	8	9	10	2	12	13	14
	15	8	9	10	7	7	13	14
	16	8	9	10	1	9	13	14
	17	8	9	10	3	11	13	14
	18	8	9	10	11	11	13	14
	19	8	9	10	12	14	13	14
	20	8	9	10	7	4	13	14

And the final (fifth) A matrix is:

		Subsequence						
		1	2	3	4	5	6	7
	1	8	9	10	13	8	13	14
	2	8	9	10	9	1	13	14
	3	8	9	10	11	11	13	14
	4	8	9	10	8	9	13	14
	5	8	9	10	11	14	13	14
S	6	8	9	10	4	8	13	14
e	7	8	9	10	3	12	13	14
q	8	8	9	10	1	12	13	14
u	9	8	9	10	3	11	13	14
e	10	8	9	10	10	6	13	14
n	11	8	9	10	12	5	13	14
c	12	8	9	10	4	5	13	14
e	13	8	9	10	7	13	13	14
	14	8	9	10	2	12	13	14
	15	8	9	10	7	7	13	14
	16	8	9	10	1	9	13	14
	17	8	9	10	3	11	13	14
	18	8	9	10	11	11	13	14
	19	8	9	10	12	14	13	14
	20	8	9	10	7	4	13	14

Of course, these subsequences start at randomly varying points in the data sequences (which the program records). Note that the algorithm has already closed in on the regularity at the end of the third pass, but it is shifted to the right, missing the 8 that begins the regularity. To speed convergence, the LEA algorithm computes at regular intervals (here, after each pass through the data) a global quality measure for "phase-shifted" A matrices. These are matrices in which the whole array moves to the candidates beginning one or two elements right or left of the current location in each data sequence. These global quality measures are then added and a new A matrix is chosen as in the sampling phase above. (Note that if one candidate is very close to the beginning or the ending of its sequence, this may forbid such phase shifts in one or the other direction.)

This algorithm thus finds what seems like a needle in a haystack. It finds a relatively short regularity, with some noise in the middle, buried at random points in long sequences, without knowing what is in the regularity or where it starts in the different sequences. In fact, experi-

ment shows that the algorithm will successfully find a regularity even if it appears in only a small number of the data sequences and has substantial noise in it.

How does this happen? It happens because of the formal structure of the algorithm. We choose the new candidate from each successive focus sequence by using the conditional likelihood of each possible candidate in that sequence given the current (fixed) values of the candidates from other sequences. By doing this for sequence after sequence, through the whole data set, again and again, we set up a probability process. This process has 44^{20} points in its state space (because the state space is the space of all possible A matrices), making it somewhat intractable to direct analysis. But at any given point in that space, we can always follow the two-phase procedure above to write a current transition rule to choose our next point. And this rule will *always* produce the same probabilities for our next choice of candidate should we accidentally revisit the same point later in the process (that is, should we revisit an earlier version of the A matrix) because that point will produce an identical Q matrix in the Q-calculation phase. Therefore the process we have created is a stationary Markov chain and has a limiting distribution. Moreover, a theorem of Geman and Geman (1984:732, theorem C) guarantees that under any scanning procedure that guarantees a finite recurrence time between successive updates of the entire A matrix, the distribution of any function of that matrix will also approach a limit.

Note that the results depend on the probabilistic nature of the process. The Gibbs sampler rates the changes it could make in a given row of the A matrix and does indeed know the "best way to go uphill in one dimension." But it does not take that direction. Rather it chooses a direction with a probability given by the current conditionals. If the best way were always taken (if the sampling phase followed a simple "take the best" rule, rather than a sampling rule), there would be no guarantee that every possible point in the state space could be visited from every possible beginning, therefore no stationary irreducible Markov chain on the entire state space, and no guarantee that the transition rule's equilibrium distribution would be that of the space. The negative practical result would be the possibility of stopping at some local maximum.[17]

_____

[17] All LEA calculations were done by an implementation of the algorithm in QUICKBASIC 4.5 on a 486 clone, written by the senior author. All detailed calculations (Bayesian priors, etc.) are implemented exactly as described in the original LEA paper. Lawrence et al. (1993) report several optimization criteria, of which the information statistic presented in text is the simplest. Note that the algorithm as imple-

The LEA algorithm, like most similar algorithms, wanders through a good deal of the data before converging. However, when random chance throws a few of the true best common subsequences into the candidate matrix, the algorithm focuses effectively. It can converge even in the presence of quite minimal regularities. Convergence is measured in various ways. The algorithm is effectively maximizing the product of the "quality quotients" across the candidates in the A matrix, and Lawrence et al. derive a variety of global statistics related to that product. As their best summary statistic, Lawrence et al. present an "information per bit" statistic, which we report here without discussion. (Higher values are better. Random data has an information statistic around 1.0. Perfect subsequences—replicated exactly in every data sequence—have information statistics around 3.0 in our experience. Unlike other measures discussed by Lawrence et al, the information statistic is not sensitive to subsequence length.)[18]

4.3. *Analysis of Article Subsequences*

In this application, we consider subsequences only of length two and three. At length four, we are already considering over half of the entire article sequence in one-third of the cases.

Applied to the present data, the LEA algorithm does converge to a best common subsequence. We found the same result in three separate runs of 25 iterations (one iteration uses each of the 99 sequences as focus se-

mented here does not permit insertion. All subsequences are true successive subsequences.

The rate of convergence of the Gibbs sampler is a matter of hot debate and grave concern. For a variety of perspectives, see the comments in the number of the *Journal of the Royal Statistical Society* mentioned in n. 12 above. See also Kong, Liu, and Wong (1994), Liu, Wong, and Kong (1994, 1995). Papers related to the statistical properties of the LEA algorithm are Liu Neuwald, and Lawrence (1995) and Neuwald, Liu, and Lawrence (1995).

In practice, the LEA algorithm converges very fast. One commenter on this exposition pointed out that in fact that the algorithm's developers do not know why it works so quickly; it's a "happy accident."

[18]The discussion of these statistics is found in Lawrence et al. (1993:209–210, 214, fn. 22). The information per bit statistic for the five passes of the example just given are

$$INFO = .8228551$$
$$INFO = .7174581$$
$$INFO = .8768407$$
$$INFO = 1.457156$$
$$INFO = 2.135363.$$

quence exactly once). That subsequence is 15–43–21 (Description of research design—Data analysis/Summary of findings—Total summary of findings). It appears in six out of the 99 sequences, those six coming from volume years 51, 59, 62, 63, 64, and 66. Thus the single strongest three-element subsequence appears in 2 of 15 articles in the next to last decade and in 4 of 17 articles in the last. Given this result, the algorithm will have found both 15–43 and 43–21 if they exist as two-step subsequences in other data sequences. But they are not particularly common. The first appears twice, in volume years 56 and 67. The second appears four times, in volume years 40, 52, 57, and 59. (Since element 43 is part of this common subsequence and appears seldom before the fifth volume decade, all articles using it are late.)

       After we deleted these six sequences from the data set, we ran the LEA algorithm again to see if there was a second best common subsequence, another version of the "rhetorical armature." There is another such common subsequence: 10–15–53 (Author's hypotheses—Description of Research Design—Data analysis/Summary of findings/Discussion of relevance of findings to theory and hypotheses). This appears in four articles, in volumes 44, 49, 62, and 69. Note that element 53 includes element 43 as a subset, and thus that this pattern is close to the first part of the more common 15–43–21 subsequence. In summary, about 10 of the 47 sampled articles in the journal's last three decades follow one of two versions of a common rhetorical subpattern. Both the first and second most common subsequences come from the latter part of the data, even though element 40—the characteristic "analysis" element of the earlier period—is more common than element 43. Indeed, the pattern 10–15–53 appeared only on the third of three runs of 25 iterations, of the reduced data set, the other two being dominated by element 40 in the second or third column in the candidate matrix. But there are not as many common patterns employing element 40 as there are employing 43 and its relative 53, and so repeated iteration ultimately escapes the clutches of element 40.

       To test whether these results were due to chance, we performed the following iteration:

1.   Shuffle the elements within each sequence.
2.   Run the LEA algorithm for 25 iterations on the resulting data.

We did steps 1 and 2 100 times to generate a baseline distribution for the information per bit statistic in this data. This can be referred to a standard

normal probability table on the assumption that the underlying distribution is normal. (Inspection of normal probability plots for the data verified this assumption. There is also a nonparametric significance test based on the Wilcoxon signed-ranks test; see Neuwald, Liu, and Lawrence 1995 and Liu, Neuwald, and Lawrence 1995.)

For the full data, the resulting information statistic distribution had mean 1.18 and standard deviation 0.16. The unshuffled value was 1.83, a z-value of 4.04. For the data with those articles possessing the most common subsequence deleted, the mean was 1.17 and the standard deviation 0.17. The unshuffled value was 1.74, a z-value of 3.40. In both cases, the results are clearly significant.

The subsequence evidence, therefore, does suggest a slight move toward rhetorical tightness. Nonetheless, it remains clear that the new model was hardly a straightjacket. Most articles did not use it. Even in the final decade, only six articles out of 17 followed one of the common subpatterns.

5. CONCLUSION

Substantively, these analyses show that the evidence for rhetorical rigidification in sociology articles is weak indeed. While there is some evidence of more recent authors' writing to a clearer template (the subsequence evidence), the actual results in both early and late periods remain diverse. It is more likely that there was a change from a fairly clear rhetorical pattern in the 1920s and 1930s to a different clear rhetorical pattern after the war. But even given the clear patterns, authors at both times produced diverse structures indeed. There is no serious evidence here of rising rhetorical hegemony. The rhetorical rigidity seen by analysts like Bazerman in modern social science is probably in the eye of the beholder.

On the methodological side, the LEA algorithm can see broad use in sociology. There are many kinds of data where we are interested in short subpatterns. Sometimes, as in the present case, these are subpatterns in fixed linear orders: rituals, texts, performances and the like. At other times, they will be portions of processes in time, like careers and life cycles. In particular, the concept of "turning point" seems ideally suited to investigation by subsequence methods. Current theory sees turning points in a variety of social processes (see Abbott 1996b for a review). Yet the LEA algorithm can be used to focus directly on such particular, crucial parts of

processes, separating them from what may well be extended periods of pure randomness or pure stability that surround them.

More broadly, Gibbs sampling and other MCMC methods will have important uses throughout sociology. Many of these uses will be in improving estimation "behind the scenes" of models that are already widely used in canned programs. But there are important substantive areas that could be revolutionized by such methods. In particular, network analysis seems a likely possibility. Gibbs sampling methods arose within the theory of Markov random fields, within which the Ising model looms large (See Geman and Geman 1984). And the Ising model is the origin of blockmodeling with its emphasis on local correlations (Breiger, personal communication). Having helped transform that area of physics, Gibbs sampling and its relatives will most likely have the same kinds of implications for network analysis. We can look forward to their broad application in sociology.

REFERENCES

Abbott, Andrew. 1992. "From Causes to Events." *Sociological Methods and Research* 20:428–55.

———. 1995a. "Sequence Analysis." *Annual Review of Sociology* 21:93–113.

———. 1995b. "On Dijkstra and Taris: The Scope of Alignment Methods." *Sociological Methods and Research* 24:232–43.

———. 1996a. "AJS: The Growth of an Institution." Unpublished paper, University of Chicago.

———. 1996b. "On the Concept of Turning Point." In *Comparative Social Research*, edited by F. Engelstad et al. Greenwich CN: JAI Press. Forthcoming.

Abbott, Andrew, and Alexandra Hrycak. 1990. "Measuring Sequence Resemblance." *American Journal of Sociology* 96:144–85.

American Journal of Sociology. 1966. *Cumulative Index to the American Journal of Sociology, 1895–1965*. Chicago: University of Chicago Press.

Assadi, B. 1987. "The Social Construction of Knowledge in American Sociology." Ph.D. diss., Howard University.

Bakanic, Von, Clark McPhail, and R. J. Simon. 1987. "The Manuscript Review and Decision-Making Process." *American Sociological Review* 52:631–42.

Barthes, Roland. 1974. *S/Z*. New York: Hill and Wang.

Bazerman, Charles. 1988. *Shaping Written Knowledge*. Madison: University of Wisconsin Press.

Boguski, Mark S. 1992. "Computational Sequence Analysis Revisited." *Journal of Lipid Research* 33:957–74.

Buehler, C., Gary Hesser, and Andrew J. Weigert. 1972. "A Study of Articles on Religion in Major Sociology Journals." *Journal for the Scientific Study of Religion* 11:165–70.

Casella, George, and Edward I. George. 1992. "Explaining the Gibbs Sampler." *American Statistician* 46:167–74.

Champion, D. J., and M. F. Morris. 1974. "A Content Analysis of Book Reviews in AJS, ASR, and Social Forces." *American Journal of Sociology* 78:1256–65.

Clyne, M. 1987. "Cultural Differences in the Organization of Academic Texts." *Journal of Pragmatics* 11:211–47.

Connors, R. J. 1985. "The Rhetoric of Explanation." *Written Communication* 2:49–72.

Cox, D. R., and H. D. Miller. 1965. *The Theory of Stochestic Processes* London: Chapman and Hall.

Everitt, Brian S. 1993. *Cluster Analysis* London: Edwin Arnold.

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

Garnett, R. A. 1988. "The Study of War." *American Sociologist* 19:270–82.

Gelfand, Alan E., Susan E. Hills, Amy Racine-Poon, and A. F. M. Smith. 1990. "Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling." *Journal of the American Statistical Association* 85:972–85.

Gelfand, Alan E., and Adrian F. M. Smith. 1990. "Sampling-based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association* 85:398–409.

Geman, Stuart, and Donald Geman. 1984. "Stochastic Relaxation, Gibbs Distributions, and Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6:721–41.

Gilks, Walter R., Sylvia Richardson, and David J. Spiegelhalter, eds. 1996. *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Glenn, Norval D. 1971. "American Sociologists' Evaluations of 63 Journals." *American Sociologist* 6:298–303.

Gotoh, O. 1993. "Optimal Alignment Between Groups of Sequences and Its Application to Multiple Sequence Alignment." *CABIOS* 9:361–70.

Gribskov, Michael R., and John Devereux, eds. 1991. *Sequence Analysis Primer*. New York: Freeman.

Gusfield, D. 1993. "Efficient Algorithms for Multiple Sequence Alignment with Guaranteed Error Bounds." *Bulletin of Mathematical Biology* 55:141–54.

Gusfield, J. 1981. *The Culture of Public Problems.* Chicago: University of Chicago Press.

Henikoff, S., and J. G. Henikoff. 1994. "Position-Based Sequence Weights." *Journal of Molecular Biology* 243:574–78.

Ishikawa, M., T. Toya, M. Hoshida, K. Nitta, A. Ogiwara, and M. Kanehisa. 1993. "Multiple Sequence Alignment by Parallel Simulated Annealing." *CABIOS* 9:267–73.

Jain, Anil K., and Richard C. Dubes. 1988. *Algorithms for Clustering Data* Englewood Cliffs, NJ: Prentice Hall.

Karlin, Samuel, P. Bucher, V. Brendel, and S. F. Altschul. 1991. "Statistical Methods and Insights for Protein and DNA Sequences." *Annual Review of Biophysics and Biophysical Chemistry* 20:175–203.

Kinloch, G. C. 1988. "American Sociology's Changing Interests as Reflected in Two Leading Journals." *American Sociologist* 19:181–94.

Kircz, J. G. 1991. "Rhetorical Structure of Scientific Articles." *Journal of Documentation* 47: 354–472.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecci. 1983. "Optimization by Simulated Annealing." *Science* 220:671–81.

Kong, Augustine, Jun S. Liu, and Wing H. Wong. 1994. "Sequential Imputations and Bayesian Missing Data Problems." *Journal of the American Statistical Association* 89:278–93.

Lawrence, Charles E., Stephen F. Altschul, Mark S. Boguski, Jun S. Liu, Andrew F. Neuwald, and John C. Wooton. 1993. "Detecting Subtle Sequence Signals." *Science* 262:208–14.

Liu, Jun S. 1991. *Correlation Structure and Convergence Rate of the Gibbs Sampler*. Unpublished Ph.D. diss., University of Chicago.

Liu, Jun S., Andrew F. Neuwald, and Charles E. Lawrence. 1995. "Bayesian Models for Local Sequence Alignment and Gibbs Sampling Strategies." *Journal of the American Statistical Association* 90:1156–70.

Liu, Jun S., W. H. Wong, and A. Kong. 1994. "Covariance Structure of the Gibbs Sampler with Applications to the Comparisons of Estimators and Augmentation Schemes." *Biometrika* 81:27–40.

———. 1995. "Covariance Structure and Convergence Rate of the Gibbs Sampler with Various Scans." *Journal of the Royal Statistical Society*, Series B, 57:157–69.

McCartney, James L. 1970. "On Being Scientific." *American Sociologist* 5:30–35.

McCloskey, Donald. 1985. *The Rhetoric of Economics*. Madison: University of Wisconsin Press.

Metropolis, N., and S. Ulam. 1949. "The Monte Carlo Method." *Journal of the American Statistical Association* 44:335–41.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. "Equation of State Calculations by Fast Computing Machines." *Journal of Chemical Physics* 21:1087–92.

Mullins, N., W. Snizek, and K. Oehler. 1988. "The Structural Analysis of a Scientific Paper." *Handbook of Quantitative Studies of Science and Technology*, edited by A. F. I. Raam, 81–105, New York: Elsevier.

Murray, S. O. 1988. "The Reception of Anthropological Work in Sociology Journals, 1922–1951." *Journal of the History of the Behavioral Sciences* 24:135–51.

Myers, G. 1990. *Writing Biology*. Madison: University of Wisconsin Press.

Needleman, S. B., and C. D. Wunsch. 1970. "A General Method Applicable to the Search for Similarities in the Amino-Acid Sequence of Two Proteins." *Journal of Molecular Biology* 48:443–53.

Neuwald, Andrew F., Jun S. Liu, and Charles E. Lawrence. 1995. "Gibbs Motif Sampling." *Protein Science* 4:1618–32.

O'Neill, J. 1981. "The Literary Production of Natural and Social Science Inquiry." *Canadian Journal of Sociology* 6:105–20.

Page, Charles H. 1981. "*The American Sociological Review*, 1958–1960." *American Sociologist* 14:43–47.

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

Smalley, T. N. 1981. "Trends in Sociology Literature and Research." *Behavioral and Social Sciences Librarian* 2:1–19.

Tanner, Martin A. 1991. *Tools for Statistical Inference*. New York: Springer.

Tanner, Martin A., and Wing H. Wong. 1987. "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Association* 82:528–40.

Teevan, J. J. 1980. "Journal Prestige and Quality of Sociological Articles." *American Sociologist* 15:109–12.

Vingron, Martin, and Michael S. Waterman. 1994. "Sequence Analysis and Penalty Choice." *Journal of Molecular Biology* 235:1–12.

Vingron, Martin, and Peter R. Sibbald. 1993. "Weighting in Sequence Space." *Proceedings of the National Academy of Science* 90:8777–81.