Toward Computational Literature Reviews:
Analyzing Theories in Complex Texts through Supervised and Unsupervised Text-Analysis Methods[1]

Jaren Haber, Heather A. Haveman, and Yinuo Xu

February 20, 2023

Abstract

Academic research requires reviewing the literature to determine what we know and don't know. But the number of journals and articles published has increased enormously, making literature reviews challenging, especially for interdisciplinary fields. In this paper, we develop a flexible and reproducible method for computational literature reviews that takes advantage of massive online article collections. As a test case, we use journal articles from JSTOR to investigate the evolution of organizational theory, an interdisciplinary field developed primarily by sociologists and management scholars. We focus on three prominent organizational theories: organizational ecology, resource dependence, and organizational institutionalism. We develop an inductive method of applying expert-built dictionaries for automated analysis of complex texts, then expanding those dictionaries using word embeddings to ensure they are comprehensive. We then measure the literature's engagement with each perspective from 1971 to 2014 in two ways: (1) applying a dictionary-based measure to the JSTOR corpus and (2) counting citations to core articles for each theory using the Web of Science. We use hierarchical clustering to assess whether theories became more consolidated (more coherent) or more fragmented (less coherent), and whether each theory became more or less distinct from the other two.

[1] We acknowledge the support of grants (SES170018 and SES180020) from the National Science Foundation for access to XSEDE (eXtreme Science and Engineering Discovery Environment) for data storage and processing capacity. We thank Aaron Culich of the D-Lab and the Computational Text Analysis Working Group at U.C. Berkeley for helpful advice and comments on this project. We also thank the following undergraduate research assistants, who worked with us through the Data Science Discovery and Undergraduate Research Apprentice Programs at U.C. Berkeley: Thomas Lu, Yoon Sung Hong, Rebecca Abraham, Zekai Fan, Rudy Venguswamy, Deepak Ragu, Laiming Huang, Shixuan Li, Jahnavi Singh, Yijun Long, and Poorvi Acharya.

Research in many academic fields requires reviewing the literature to document what we know about a phenomenon and what gaps exist in our knowledge. For interdisciplinary fields, literature reviews are especially challenging because there are more, and more varied, publication outlets, ranging from blogs to working papers in online archives, conference proceedings, journal articles, chapters in edited books, and entire books. Even if we restrict attention to articles published in peer-reviewed journals, the primary outlet for many scientific fields, the number of journals has grown exponentially, doubling every 10 to 15 years (Price 1965; Mabe and Amin 2001). Moreover, recent years have seen an efflorescence of online-only journals such as the *Public Library of Science* (*PLoS*). This explosion of journals makes surveying the field of scholarly knowledge increasingly difficult and prone to sample-selection bias.

In this paper, we offer one solution to the problem of reviewing such vast literatures: a computational workflow for reviewing academic literature that takes advantage of online collections containing nearly all articles published in academic journals (e.g., JSTOR, MathSciNet, Web of Science, MEDLINE). Our computational review stands in sharp contrast to the usual practice of reading a small number of texts, usually far less than 1,000. Human reading is prone to selection and evaluation biases stemming from researchers' training and social networks, so it has low inter-rater reliability. In contrast, computational methods can (within limits) code the same text the same way every time. They can generate reliable literature reviews that have the potential to accelerate research. Finally, we have made our code publicly available, which ensures that our methods are reproducible.

As our case study, we review the literature in organizational theory, an interdisciplinary field that was developed primarily by sociologists and management scholars and secondarily by scholars in many other social science departments (economics, history, political science, and psychology) and professional schools (education, engineering, labor relations, law, and public policy). We focus on three prominent organizational theories: organizational ecology, resource dependence, and organizational institutionalism. These theories invoke different concepts and propose different relationships among concepts.

We trace how extensively these three theories are used in the literature in sociology and management (the fields in which most organizations research is conducted) using journal articles from JSTOR, a repository that has excellent coverage of the social sciences and business fields. We begin by drawing on expert knowledge to generate theoretically distinctive dictionaries of terms (words and phrases) from a set of foundational texts for the three theories. We expand these dictionaries using word embeddings to discover semantically similar terms based on linguistic trends in specific time periods. We then use our expanded dictionaries to measure journal articles' engagement with each theory over time through word counts. Finally, we use similarity metrics and hierarchical clustering to assess when and to what extent the concepts invoked in each theory became more or less linked in the literature.

Over the past decade, several scholars have speculated about which organizational theories are thriving or dying (e.g., Davis 2010; Alvesson and Spicer 2019). Our analysis shows that all three remain thriving, although they are not all used equally. As time passed, journal articles in sociology and management engaged far more with concepts from institutionalism and resource dependence than from organizational ecology.

The paper proceeds as follows. In the next section, we explain the benefit of focusing on article content (rather than, for example, citation patterns) to explain the extent to which a scholarly perspective is used. After that, we describe our research methods, including samples, measures, and analytical techniques. Then we present the results of our analyses and outline our contributions.

Theory

Language reflects culture: the cognitive categories through which people attend to the world are embedded in the words they use (Whorf 1956; Sapir 1958). Words that are used frequently are cognitively central and reflect what is most on the speaker's or writer's mind. Words that are used infrequently or not at all are at the speaker's cognitive periphery. Infrequently used words may even represent uncomfortable or alien concepts.

One common way language reflects culture is through specialized vocabulary, words and phrases coined by cultural subgroups to capture important ideas that are important to them – in other words, jargon. Jargon makes it possible to communicate key ideas quickly and precisely among the members of cultural subgroups, including scientists. As French philosopher Étienne Bonnot de Condillac asserted, "Every science requires a special language, because every science has its own ideas" (de Condillac 1782; quoted in Braudel 1982 [1992]: 234). Every scientific community possesses a long list of concepts that are central to its theories – these are de Condillac's "ideas." Through their use of jargon, scholars align their writing with their particular intellectual community (Vilhena et al. 2014), with their own epistemic culture (Knorr-Cetina 1999). For example, when Marxist scholars discuss *domination* and *exploitation*, those terms highlight the importance of unequal power between workers and bosses. Similarly, when social-exchange theorists discuss *reciprocity* and *reputation*, these words encapsulate the idea that all exchanges are affected by the possibility of future exchanges and the results of past exchanges. And when family demographers discuss *marriage and fertility rates*, these phrases are efficient shorthand for common life-changing events that vary greatly across regions, types of people, and over time. The lists of concepts developed by different scientific communities are distinctive – they overlap little, if at all. For example, social-exchange scholars are far less likely to mention domination or fertility rates than they are to talk about reputation. Accordingly, to capture the theoretical concepts that are important to different intellectual communities, we must focus on the specialized vocabulary used by each community. For us, the intellectual communities are the organizational theorists who use the three theories we study.

Case Study

To illustrate our computational method for reviewing scholarly literatures, we need a test case. Here, we use the literature on organizations, which studies how and why organizations are formed, how they operate, and how they interact (Haveman 2022). The study of organizations has a long history, dating back to the nineteenth century. It spans multiple social science disciplines (e.g., anthropology, sociology, political science) and professional schools (e.g., business, engineering, law), although it is largely conducted by sociologists and management scholars. It also spans many levels of analysis, from very micro, meaning individual members of organizations (e.g., employees in firms, patients in hospitals), to very macro, meaning fields of interacting organizations (e.g., the health-care field).

Despite its variety, the organizations literature has been dominated since the 1970s, by three perspectives – demographic, relational, and cultural – that map onto general sociological conceptions of social structure and agency (Haveman 2022). The *demographic perspective* holds that social structure is constituted by distributions of organizations along salient dimensions of social and physical space, such as strategy, size, and location. The *relational perspective* holds that social structure is constituted by webs of social connections among organizations; e.g., ownership or buyer-supplier ties. The *cultural perspective* holds that social structure is constituted by widely shared norms, values, expectations, roles, and rituals, which generate understandings of what is possible and reasonable. These three perspectives have different conceptions of organizational identity motivations for action. For demographers, identity and motivation derive from position, absolute or relative, along salient dimensions of social life. For relational scholars, identity and motivation derive from ties among organizations. For cultural scholars, identity and motivation derive from social interaction.

To make our analysis manageable, we focus on one theory in each perspective: organizational ecology (from the demographic perspective), resource-dependence theory (relational), and organizational institutionalism (cultural).[2] We chose these theories for three reasons: they are prominent, they emerged around the same time (in the mid 1970s), and they examine the same units of analysis (entire organizations and industries). These three organizational theories highlight different concepts and relationships between those concepts. In the interests of space, we discuss only a few, shown in italics below, to illustrate the differences between the theories; for a more complete description, see Haveman (2022). *Organizational ecology* assumes that organizations are *structurally inert*: it is difficult for them to *change* their strategies and structures because they face pressures for *reliability* (consistent performance) and *accountability*, and change can harm performance and increase the chance of *failure* (Hannan and Freeman 1977, 1984). The rise of *vested interests*, *limits on information*, *barriers to entry and exit* from market niches, and *legitimacy* concerns all contribute to structural inertia. In contrast, *resource-dependence theory* focuses on *power-dependence relations* and how dependence makes organizations vulnerable to *influence* attempts (Pfeffer and Salancik 1978; Burt 1983). It also predicts a wide array of responses, including finding *alternative exchange partners*, *merging*, or forming *alliances* or *joint ventures*. Finally, *organizational institutionalism* focuses on how and why organizations *conform* to *institutionalized rules* in order to achieve *legitimacy* (Meyer and Rowan 1977). Legitimacy, in turn, is assumed to make it easier for organizations to acquire *resources* and *survive*. Legitimacy can be achieved by *decoupling* formal structures (what organizations say they do) from on-the-ground practices (what organizations actually do). Legitimacy can also be achieved if organizations become institutionally *isomorphic* (similar) to other organizations due to

[2] Each of the three perspectives on organizations (demographic, relational, and cultural) contains many theories. The demographic perspective includes organizational ecology and theories of internal organizational demography. The relational perspective includes resource-dependence theory, theories of social capital, and theories of inter- and intra-organizational network structure. And the cultural perspective includes theories of organizational culture and organizational institutionalism. The perspectives are not mutually exclusive: any research product (book, book chapter, or paper) could combine any two of the perspectives, or all three.

coercive, *normative*, or *regulative* pressures (DiMaggio and Powell 1983). Although these three theories all take as their unit of analysis the organization or industry, they make different assumptions, focus on different causal mechanisms, and predict different outcomes. The concepts they invoke overlap only a little; e.g., legitimacy is part of both ecology and institutionalism.

Research Methods

Sampling Plan and Data Preparation

To understand how the influence of these three organizational theories has evolved over time – how much published literature has engaged with them conceptually – we conducted a computational literature review. We focused on articles published in peer-reviewed journals because organizational theory research largely appears in articles rather than books. Moreover, many influential organizational-theory books (e.g., Pfeffer and Salancik 1978; Burt 1983; Hannan and Freeman 1989; Powell and DiMaggio 1991) repackage material that was first published in article form.

Our main data are published articles from JSTOR, a repository that has excellent coverage of both fields where organizational theory research articles appear: sociology and management.[3] As of November 2018, JSTOR listed 38 journals in the subject "management and organizational behavior" (hereafter referred to as management) and 150 journals in the subject "sociology." Given these numbers of journals, no scholar can survey all the published literature. Our computational approach allows us to analyze the full population of published articles, rather than a small sample, to sift out articles that are not on organizations, and classify articles on organizations by their engagement with the three theoretical perspectives.

We received data from JSTOR covering the years 1971 to 2014 (inclusive) in July, 2018.[4] These data contained all texts published in journals that were in the two JSTOR subject areas central to organizational theory: sociology and management (N = 398,703). The Technical Appendix provides more details about the raw data. Although JSTOR does not include all peer-reviewed scholarly journals in sociology and management – publishers have to compensate JSTOR (a non-profit organization) to be included – it does cover the most prestigious journals and many less well known journals.

As we explain in the Technical Appendix, we filtered out 82,948 texts published in journals whose primary subject area was neither sociology nor management, as well as 160,202 texts in journals published in languages other than English.[5] That left 155,553 texts published in 38 journals in management and 78 journals in sociology. Our focus is on full-length scholarly articles, so we excluded from analysis 85,471 book reviews, lists of books received, front matter (e.g., tables of contents), along with 4,717 articles shorter than 1,000 words (e.g., errata).

[3] We received the data from JSTOR's Data for Research group. This service is now provided through Constellate, a project of JSTOR Labs: https://about.jstor.org/whats-in-jstor/text-mining-support/.

[4] Articles from years more recent were not available at the time of our request due to the waiting period included in the agreements JSTOR has with many academic journals. Such agreements allow journal publishers to generate subscription income from researchers and libraries.

[5] A tiny number of journals that publish articles in multiple languages, including English, were excluded (e.g., *Acta Historica Academiae Scientarium Hungaricae* and *Anabases*).

Given our focal interest in tracking the prevalence and shape of organizational theories over time, we also removed 2,327 articles missing data on year of publication. Finally, since our focus is on organizational studies, we also filtered out any articles that did not mention at least one word common in that lexicon, which were mostly synonyms for the word 'organization' or its components (e.g., 'department').[6] This linguistic filter removed 3,940 articles, including 2,736 from sociology and 1,204 from management. Our final dataset included 59,098 articles, with 15,008 in management and 44,090 in sociology.

Analytical Methods

Our analysis began by developing dictionaries for the three theories, then used an unsupervised machine-learning technique, word embeddings, to inductively add terms that were closely related to the original dictionaries. For this, we used the word2vec algorithm (Mikolov, Sutskever, et al. 2013), the most commonly used word-embedding algorithm in social science research. We counted terms from these expanded dictionaries in our corpus to measure each journal article's engagement with each theory, and we validated this by counting citations to foundational articles in each theory in an external database (Web of Science). Next, we used hierarchical clustering and similarity metrics to capture theoretical consolidation among and distinctiveness between the three theories over time. While the larger time scale of this method reveals coarser time trends than year-over-year measures such as engagement, representing our concepts as word vectors allows us to capture broader linguistic changes in the consolidation of our theories and their time-varying expression through specific words. We describe each step in turn below.

Creating initial (seed) dictionaries. Our analysis is based on dictionary methods, rather than citation-based methods (e.g., Moody and Light 2006; Vogel 2012), semantic network analysis (e.g., Vilhena et al. 2014), or topic modeling (e.g., Hall, Jurafsky, and Manning 2008). Dictionary methods represent documents by counting whether and how much texts include the terms (words and phrases) in dictionaries (Stone et al. 1966). They are appropriate when categories (here, theories) and textual features (here, the words and phrases associated with each theory) are known and fixed (Quinn et al. 2010). Dictionary methods have been used extensively in the social sciences, not just in sociology (e.g., Goldberg et al. 2016), but also in political science (e.g., Laver and Garry 2000), finance (e.g., Tetlock 2007), and psychology (e.g., Snefjella and Kuperman 2015). To apply dictionary methods, researchers develop lists of terms connected to different categories of interest, then categorize texts by counting instances of these terms (Grimmer and Stewart 2013; Jurafsky and Martin 2023). Dictionaries can be hand-driven (Schwartz and Ungar 2015); that is, developed by experts from domain knowledge and/or identifying synonyms and syntactic variants through tools like WordNet (Miller 1985). Or they can be data-driven; that is, developed by crowd-sourcing (e.g., Dodds et al. 2011; Benoit et al. 2016) or identifying distinguishing words from labeled corpora (Monroe et al. 2008).

[6] The full list of terms we used for this filter is as follows: 'organization', 'organizational', 'organizations', 'firm', 'firms', 'association', 'associations', 'employer', 'employing', 'employment', 'bureaucracy', 'bureaucracies', 'bureaucratic', 'office', 'offices', 'bureau', 'bureaus', 'department', 'departments', 'departmental', 'subunit', 'subunits'.

Most dictionary-based analysis uses validated, publicly available dictionaries such as the Linguistic Inquiry Word Count dictionaries (Pennebaker et al. 2007). We are interested in something very different, specifically theories of the middle range (Merton 1968), which contain complex sets of concepts (terms consisting of individual words and multi-word phrases) that are logically interrelated. To capture such complex sets of terms, we followed other research on complex concepts (e.g., Graham, Haidt, and Nosek [2009] on political ideology; Bartels, Oliver, and Rahn [2016] on anti-establishment rhetoric) and developed three custom dictionaries, each containing core concepts in one of the three theories we study. For this, we used the second author's expert knowledge of organizational theories and their associated concepts. She began by reading the foundational texts of each perspective. For organizational ecology, she used Hannan and Freeman (1977, 1984), the foundational theoretical articles. For resource-dependence theory, she used Pfeffer and Salancik (1978). She focused on chapters 1-2 and 5-8 of this book because they are about power and relationships among organizations, rather than within organizations. She examined the theoretical sections, not the reports of empirical analyses. And for organizational institutionalism, she examined the two foundational articles, Meyer and Rowan (1977) and DiMaggio and Powell (1983).

From these texts, the second author recorded key terms (individual words and two-word phrases) and their cognates (e.g., imitate, imitation, imitating) and removed the same stop words as were removed during the creation of the JSTOR data (see the Technical Appendix). She then compared the three dictionaries and eliminated overlapping terms. For example, she deleted "organization," "uncertain," and "uncertainty" from all three dictionaries; "homogeneous" from the ecology dictionary because it was more central to institutionalism; "relational network" from institutionalism dictionary because it was more central to resource dependence; and "survival" from the resource-dependence dictionary because it was more central to ecology. The only overlapping terms she retained were "legitimate" and its cognates in ecology and institutionalism dictionaries because these concepts were central to both theories. Finally, she removed compound terms with shared roots (e.g., "ritual commitment" and "ritual confidence" were reduced to the word "ritual"). That left sharp, distinctive lists of terms.

Extending seed dictionaries using word embeddings. To ensure that the dictionaries containing terms denoting the three theories' concepts were complete and not biased by the second author's training and social networks, we followed previous research (e.g., Garten et al. 2018; Sivak and Smirnov 2019) and used a modified form of computational grounded theory (Nelson, Burk, Knudsen, and McCall 2021). Specifically, we expanded our seed dictionaries using *word embeddings*. Word embeddings map words onto high-dimensional vector spaces (typically 100-300 dimensions) and represent semantic relations between words as geometric relations in vector space (Mikolov, Sutskever, et al. 2013). Word-embedding models are a form of distributed representation: each word is "understood as the sum of all its environments," its co-occurrents (Harris 1954: 146). Words that share many contexts (i.e., they are collocated with the same other words) are positioned near each other in vector space, and words that have very different contexts (i.e., they are collocated with different other words) are positioned far apart. Word-embedding models efficiently and accurately predict semantic similarity; i.e., the extent to which any two words in a dataset are used in similar contexts and, thus, have similar meanings. Importantly, this relational mapping captures commonalities in words' local contexts, rather than collocation alone. Mapping contexts encodes word embeddings with underlying cultural meanings, rather than strictly empirically observable patterns.

Word-embedding techniques are frequently used in digital humanities research; for example, to map the geography of syntactical variants, such as adverbial intensifiers ("hella" in Oakland, CA versus "wicked" in New England; Bamman, Dyer, and Smith 2014), or to trace the shifting meaning of terms over time (e.g., over the twentieth century, "gay" moved from a position near "dapper" and "cheerful" to one close to "lesbian" and "homosexual"; Kulkarni, Perozzi, Al-Rfou, and Skiena 2015). Use of word embeddings is becoming more common in the social sciences; for example, to analyze associations between basic cultural categories such as gender, race, and status (Kozlowski et al. 2019) or to tracing changes in the use of theoretical concepts over time and comparing use across texts (Stoltz and Taylor 2019).

To train word embeddings on our corpus of academic articles, we use the word2vec (Mikolov, Sutskever, et al. 2013) implementation in the gensim library in Python (Řehůřek and Sojka 2010). We trained the word2vec algorithm on our data because custom-trained algorithms yield better results than pre-trained algorithms when the concepts under study are complex or arcane (Garten et al. 2018). To do this, we used the full corpus of articles (after removing HTML tags and author surnames). We then divided the corpus into four 11-year periods (1971-81, 1982-92, 1993-2003, 2004-14), which allowed us to trace changes in how theoretical concepts (and which concepts) were used over time as the scholarly literature evolved.

Word2vec learns associations between words and their contexts using windows of (typically 6 to 12) words on either side of focal words. Computationally, word2vec learns word vectors using a single-layer neural-network architecture (Turian, Ratinov, and Bengio 2010). For a series of words w_1, w_2, w_3, … , w_T, the goal is to maximize the average log probability of predicting w_{t+j} given w_t (Mikolov, Sutskever, et al. 2013: 2):

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p\left(w_{t+j}\middle|w_t\right)$$

where *w* indicates a word in the sentence, *t* is the iterator over *T* training words in the sample, *c* is the window size, and *j* is a number between -*c* and *c* (excluding zero) that indicates the distance from focal word w_t to a word within its context, w_{t+c}. The Technical Appendix provides more details on how we implemented the word2vec algorithm. Because the significance of any set of words shifts across linguistic contexts (Louwerse 2004) and sociocultural contexts (Henrich et al. 2010), we trained a different word-embedding model for each 11-year period.

Word-embedding models calculate distances between terms in semantic spaces that typically have 100-300 dimensions. Given the high number of dimensions, Euclidean (straight-line) distances are inappropriate for comparing semantic locations. Instead, analysts use cosine similarity to measure distances between word vectors; i.e., by calculating the cosine of the angle between two word vectors (Jurafsky and Martin 2023). This measure ranges from 1 to -1. A score of 1 means that words have the same meaning (the angle between their word vectors is 0°, i.e. they are perfectly parallel). A score of 0 means that they have orthogonal or independent meanings (the angle is 90°) because they are used in very different ways. A score of -1 means that they have opposite meanings (the angle is 180°).

We generated lists of 200 terms with strong semantic associations to the seed-dictionary terms using cosine similarity. Although high cosine similarity scores can indicate semantically similar terms, they also indicate terms that are not semantically similar but do tend to be found with the same surrounding words. We searched through the lists and retained only those terms that were similar in meaning to the seed-dictionary terms, and thus specifically related to the three theories. We ended up with lists of 50 terms for each theory in each time period. Table 1 shows the expanded dictionaries for the last time period (2004-2014). (Expanded dictionaries for the three earlier time periods are available from the first author.)

[Table 1 about here]

Measuring articles' engagement with the three organizational theories. For each theory, we measured how much each article used the focal theory's concepts, a construct we called *engagement*. In doing this, we are assuming that the more a focal article refers to the concepts in a focal theory's dictionary, the more it engages that theory intellectually. For example, if an article contains terms like accountability, structural inertia, and failure, it engages organizational ecology. But if it contains terms like interdependence, interlocking directorates, and joint ventures, it engages resource dependence. And if it contains terms like conformity, normative, and decoupling, it engages institutionalism. To measure engagement, we aggregated data on articles to the subject area (sociology or management) and year (1971 to 2014, inclusive) and used the expanded dictionary for the focal theory in the time period containing the focal year. This yielded six measures of engagement – organizational ecology sociology, organizational ecology management, resource dependence sociology, resource dependence management, organizational institutionalism sociology, and organizational institutionalism management – that varied annually. These measures are intuitive: the more times the articles published in a field in a particular year use the concepts found in a particular theory, the more likely they are to be engaged logically with that theory's core ideas.

Engagement with theory *p* by articles published in subject *s* at time *t* is defined as follows:

$$Engagement_{pst} = \sum_{a=1}^{n} \frac{engagement_{psat}}{n},$$

where the numerator is engagement by article *a* (published at time *t* in subject *s*) with theory *p*, *n* is the number of articles published during time *t* in subject *s*, and t spans a single calendar year. Note that $engagement_{pst}$ is the arithmetic average of $engagement_{psat}$ across all articles published in subject *s* during year *t*. $Engagement_{pst}$ is a continuous variable with a theoretical range of [0,1]. Zero indicates that the articles in the focal subject area in the focal year use *none* of the focal theory's concepts (at least none found in the expanded dictionary for the relevant time period) and 1 indicates that *all* articles use *all* of the theory's concepts. It equals 1 only when *every word* in *every article* in that subject published at that time is part of the theory's concept list – the expanded dictionary. However, no text – not even those we used to create the dictionary for each theory will yield an engagement score of 1 because each dictionary captures only the concepts most relevant to that theory and ignores concepts that are common across theories, such as "organization" and "uncertain," as well as concepts that are common in sociological and management research, such as "variable" and "theory."

Coherence and distinctiveness of theories. We use the term *coherence* to refer to the average cosine similarity of all word vectors in a dictionary, reflecting the degree of theoretical tightness or centeredness in the vector 'cloud' representing a theory in the embedding space. In contrast, *distinctiveness* here refers to the average cosine distance (the inverse of cosine similarity) between two dictionaries, indicating the degree of separation between the respective 'clouds' representing each theory. Clustering approaches like hierarchical clustering are built on such comparisons, aiming to maximize both within-cluster similarity (coherence) and across-cluster distance (distinctiveness). In addition to creating hierarchical cluster models (see below), we use these basic metrics of coherence and distinctiveness to reveal broad movements in the theoretical diffusion and cross-pollination of our theories over time.

Hierarchical clustering. We performed hierarchical clustering using the period-specific word-embedding models to investigate how each theory's dictionary evolved: whether each theory consolidated and became more coherent, or fragmented and became less coherent. We also assessed whether each theory became more or less distinct from the other two. While "divisive" methods like the popular *k*-means clustering method separate an initially large, singular cluster into smaller clusters, hierarchical clustering does the opposite: it begins with each sample or unit in its own cluster (called "leaves") and iteratively joins them together until they form a single cluster (the "root"). As with other clustering methods, hierarchical clusters are formed in a way that jointly minimizes the distances between data points within each cluster and maximizes the distances between clusters. A key advantage of hierarchical clustering is that it derives the number of clusters inductively, unlike divisive methods that require the user to specify (often arbitrarily) the number of clusters into which to split the samples. A key advantage of hierarchical clustering is that it derives the number of clusters inductively, unlike divisive methods that require the user to specify (often arbitrarily) the number of clusters into which to split the samples.

We perform hierarchical clustering on all unique terms in the expanded dictionaries, with each corresponding word vector initially assigned its own cluster. Dictionary terms are grouped together progressively based on their distances from one another. This provides several important insights into the relationships among the dictionary terms and thus the maturity of the three theories over time. We focus on three patterns. First, the fewer the clusters observed in a given time period, the more organized the clusters are and the greater the overall consolidation (across all three theories) in that time period. Second, the lower the average distances among articles in a given time period, the greater the overall theoretical consolidation in that time period. Third, the more terms derived from the same theory (rather than other theories) share clusters—especially when such clusters are large—the greater the consolidation of that theory in that time period.

We use scikit-learn (Pedregosa et al. 2011) to perform hierarchical clustering (with the AgglomerativeClustering method) and SciPy (Virtanen et al. 2020) to visualize the resulting models as dendrograms (tree structures). The key modeling decisions for hierarchical cluster analysis include the metric used to compute distances between samples (sometimes called "affinity") and the strategy for joining clusters (also called the "linkage criterion"). As discussed above, we use cosine similarities to measure distances between word vectors by means of their angle. Regarding the linkage strategy, we use the average distance between clusters to promote overall coherent, balanced clusters in terms of size and structure (Greenacre 2007).

Given the tree structure of hierarchical clustering models, an important choice is what threshold or cut-off to use to divide the lower portion (smaller clusters) and upper portion (larger clusters) of the resulting dendrogram, thus determining the number of clusters observed (Greenacre 2007). Rather than imposing a fixed, arbitrary threshold for clusters to merge, we based the threshold on the overall similarity between samples (this is the default method in SciPy).[7] One advantage of this scaling threshold is that it prevents the predictable decrease in the number of clusters that an arbitrary threshold would allow – in other words, it provides a more stringent test for theoretical consolidation over time.

Reproducibility. To ensure reproducibility, we analyzed data in Jupyter notebooks in the flexible, open-source Python 3 scripting language. We used Jetstream, a computing resource funded by the U.S. government; access was provided by grants from the Extreme Science and Engineering Discovery Environment (XSEDE; Towns et al. 2014).[8] We documented all scripts extensively and made them freely available online through two GitHub repositories (https://github.com/h2researchgroup/dictionary_methods and https://github.com/h2researchgroup/embeddings).

Validation. As the use of dictionary methods, especially custom dictionaries, has expanded in the past decade, social scientists have called for dictionary validation (e.g., Grimmer and Stewart 2013; Nelson et al. 2021). Answering this call, we validated the results of this dictionary-based method by hand-coding about 700 articles to provide ground truth for the presence of each perspective.

Results

We first discuss the prevalence of each theory over time using engagement and citations of canonical texts, which provides nuanced year-over-year evidence of the diffusion of our organizational theories. Then we analyze theoretical consolidation in our expanded dictionaries by looking at the coherence/distinctiveness and hierarchical clustering of their word vectors over 11-year time periods—each with its own word embedding space constructed from our data—from 1971 to 2014.

Prevalence of each perspective over time

Figure 1 traces engagement with each theory over time, based on the use of key concepts in journal articles. It has two panels: Figure 1a shows engagement in articles published in sociology journals; Figure 1b shows engagement in articles published in management journals. In both panels, the dashed green line indicates engagement with concepts in organizational ecology; the dotted blue line, resource dependence; and the solid red line, organizational institutionalism.

[7] The formula to determine the threshold in SciPy is 0.7*max(linkage_matrix), where linkage_matrix contains the cosine distances between all samples. In other words, this method scales the threshold to be 0.7 times the maximum distance between all samples. Therefore, as the samples become generally more similar – e.g., because their usage patterns are more consistent and related to one another – this threshold decreases, providing a higher bar that clusters must pass to merge.

[8] In particular, we used the Jetstream2 resource at Indiana University through allocation TG-SES180020.

[Figure 1 about here]

Figure 1a shows that engagement with concepts in organizational institutionalism in sociology journal articles was initially highest, engagement with resource dependence was intermediate, and engagement with organizational ecology was lowest. Engagement with concepts in organizational institutionalism and resource dependence generally tracked each other: sometimes one theory had more engagement, sometimes the other. Engagement with concepts in organizational ecology was always lower than the other two theories.

Figure 1b shows that engagement with concepts in resource dependence in management journal articles was initially highest, engagement with organizational institutionalism was intermediate, and engagement with organizational ecology was lowest. Engagement with concepts in resource dependence was persistently higher than with the other two theories. And in most years, engagement with concepts in organizational institutionalism was greater than with concepts in organizational ecology. After 2010, engagement with organizational institutionalism rose dramatically, almost reaching the level of resource dependence.

To contrast our word-based measure of theoretical engagement, we also counted citations to foundational articles for each theory over time using Web of Science (https://www.webofscience.com/wos/; see Technical Appendix for the specific lists of articles). Because these citation counts cover all fields of research, not just our focal disciplines of sociology and management, these provide broader evidence of the diffusion of each theory through all academic literatures. These counts are plotted in Figure 2.[9]

[Figure 2 about here]

All foundational articles for three theories were cited more over time. In the mid 1970s, the foundational articles for resource-dependence theory (dotted blue line) were cited a little more than those for the other two theories. By 1981, organizational institutionalism (solid red line) had overtaken resource dependence (dotted blue line), 45 citations to 44. Organizational ecology (dashed green line) was slower to be cited than the other two theories. But by the mid 1980s, its foundational articles were also cited more than those of resource-dependence theory: 61 citations to 30. Overall, the citation race was overwhelmingly won by organizational institutionalism. By 2000, its foundational articles had garnered 3,771 citations, compared to 2,117 for the foundational articles of organizational ecology and 1,357 for the foundational articles of resource-dependence theory. The gap in citation counts widened after that, with organizational institutionalism reaching over 1,000 citations in 2007, compared to 306 for organizational ecology and 117 for resource dependence. These trends are largely consistent regardless of the particular articles included in the cited set for each theory (for a visualization using 30 distinctive articles per theory, see the Technical Appendix).

These broader citation counts illustrate that the use of resource dependence theory may not be as widespread as our engagement results above suggest. While our word counts show that resource dependence theory is commonly used in sociology and especially management, references to its canon appear relatively rare outside these disciplinary confines.

[9] Although we have Web of Science data up to the end of 2022, we limited this graph to the years 1971-2014 in order to match Figure 1.

Rather, organizational institutionalism clearly wins the race for general influence (in terms of being cited), with organizational ecology running at an increasingly distant second place.

However, we caution that citing a work does not equate to applying and developing it, as influential works are often cited for ceremonial and status-signaling purposes. As such, while these citation trends provide a meaningful comparison, we place greater weight on our text-based, discipline-specific measures for purposes of tracking how our theories evolve over time.

Coherence and distinctiveness of theories over time

Figure 3 shows how the coherence (Figure 3a) and distinctiveness (Figure 3b) of the three theories evolved over time. As with other visuals, here we use dashed green lines for organizational ecology, dotted blue lines for resource dependence, and solid red lines for organizational institutionalism.

[Figure 3 about here]

Coherence scores generally rose slightly for all three theories, from an average of 0.209 in 1971-1981 to an average of 0.278 in 2004-2014. Distinctiveness scores declined slightly for all three theories, from an average of 0.871 in 1971-1981 to an average of 0.822 in 2004-2014. The only exception is a decline in theoretical coherence for organizational ecology between 1993-2003 and 2004-2014. This decline may be due to a new theoretical offshoot of ecology that analyzes the evolution of organizational populations—focusing on organizations' vital rates, as is traditional for that theory—and explains these with the concept of organizational forms as socially coded identity categories (Hsu and Hannan 2005; Hannan, Pólos, and Carroll 2007).

Clustering word vectors over time

Figures 4a-4d show the dendrograms of expanded dictionary terms for each theory. Figure 4a covers the first time period, 1971-1981; Figure 4b, 1982-1992; Figure 4c, 1993-2003; and Figure 4d, 2004-2014. In each figure, the y-axis shows individual words organized into clusters based on their word-embedding similarity, while the x-axis shows the linkage distance between individual words across levels of the tree. The right end of the x-axis represents a single cluster, i.e. the root of the tree. Individual words are colored by their association with the organizational theories: terms in green are distinctive of organizational ecology, blue of resource dependence, and red of organizational institutionalism. Gray indicates terms that occur in multiple dictionaries. The vertical dashed line indicates the threshold for separating clusters (discussed above); word clusters are indicated by tree branches of a shared color.

[Figure 4 about here]

Contrary to our expectations, the number of clusters observed in Figure 4 (branches that share a color) is relatively flat over time. While the 14 clusters evident in 1971-1981 appear to consolidate into 11 and 12 clusters in 1982-1992 and 1993-2003, respectively, the cluster count climbs back up to 15 in 2004-2014. Moreover, terms of shared color largely "stick together" in sizable clusters from 1971-1981 onward, suggesting theoretical consolidation from the outset.[10]

[10] This consistent differentiation of theories via shared clusters may partly be an artifact of our decision, earlier in the research pipeline, to expand the seed dictionaries by identifying concepts that most

On the other hand, the cluster merging threshold (the vertical dashed line) generally declines over time from 0.536 in 1971-1981, to 0.433 in 1982-1992, to 0.400 and 0.405 in 1993-2003 and 2004-2014, respectively. While coarser visual observation of terms' colors or cluster counts don't capture this granular change, the lower threshold suggests that all in all, the clusters representing our organizational theories become more differentiated over time. Based on this metric, most theoretical consolidation occurs between 1971-1981 and 1982-1992, given that the threshold drops 0.103 (in cosine similarity) between these periods. Moreover, given the miniscule shift in threshold of 0.005 between 1993-2003 and 2004-2014, the distinction between theories appears largely settled before 2004-2014.

Discussion and Conclusion

Our method for computational literature review aims to provide an adaptable framework for refining expert-based dictionaries with vector space models to track the evolution of scholarly theories over time. We respond to the push among social scientists to develop and implement tools for learning from massive troves of culturally relevant text data (e.g., Bail 2014; Bonikowski and Nelson 2022). In providing a framework for computer-assisted, at-scale reading of academic literature, we provide tools for hypothesis testing using specific theories as well as new discoveries via period-specific dictionary refinement and inquiry. We hope this blend of deduction and induction renders our approach flexible enough to benefit a wide range of scholars—especially those that span disciplines.

To summarize our case study, our complementary methods suggest that organizational ecology, resource dependence, and organizational institutionalism each remain the center of its own vibrant domain of intellectual inquiry, and that these are increasingly interconnected. We find that resource dependence is often deployed in the field of management, and organizational institutionalism is frequently discussed in sociology (as is resource dependence). Furthermore, while organizational ecology is the most cited of the three in the 1970s and remains relevant to theorizing in both disciplines, skyrocketing citations to organizational institutionalism in the 2000s make this theory arguably the most widespread (at least in sociology).

Our cluster models show that component concepts in all three theories cohere in clusters of meaning that differentiate as early as 1971-981 and continue to consolidate further until 1993-2003. Moreover, our cluster models and coherence measure suggest the greatest increase in theoretical consolidation between 1971-1981 and 1982-1992, which largely settled by the end of 1993-2003. Moreover, these theories also become increasingly intertwined over time, as indicated by declining distinctiveness between all three perspectives through 1993-2003. This suggests that over time, more and more organizational theorists used two or all three theories in their articles, treating the three theories as complements rather than

embody that theory in a given time period. Such dictionary expansion provides a sharper, more period-specific linguistic signal of a given theory, but this intentional refinement for similarity likely also heightens the extent to which the refined dictionaries "stick together" in the vector space and cluster models. Indeed, clustering on the seed dictionaries suggests less overall coherence of theoretical terms, but just as much overall theoretical consolidation over time, in this case observed as clusters decreasing in count from 22 in 1971-1981 to 7 in 2004-2014. (Full dendrograms of seed dictionaries are available from the first author.)

competitors. In sum, we find our theories develop quickly in the 1970s and stabilize by the 2000s, with a transitional period of gradual growth and interpenetration during the 1980s and 1990s.

Future research

While we use word embeddings for domain-specific expansion of seed dictionaries, future studies could optimize the initial dictionary development stage by drawing on additional subject-area experts as well. Given a set of foundational texts, these experts could create from scratch their own lists of theoretically distinctive terms. Alternatively, they could be delivered an already-processed list and be asked to rank, filter, and append to it to maximize theoretical fidelity. The resulting lists could be compared and integrated to validate and refine the seed dictionaries prior to refinement by inductive means (e.g., with vector space models).

Moreover, while we use external citation data as a contrast to our internal word counts, we think that citations could be connected to theoretical engagement more meaningfully by counting citations within the text data itself. Future scholars should consider using machine learning to develop methods for harvesting citation networks from heterogeneous reference styles in complex text datasets like ours. Such citation networks could be used to expand lists of foundational texts for each theory in a way similar to how we expand seed dictionaries in this study. These citations could then be linked to theoretical concepts and tracked over time.

Finally, while we have released our code to the public,[11] we aim also to provide a well-documented replication repository that includes step-by-step instructions for downloading original text datasets from JSTOR (and their new platform, Constellate: https://labs.jstor.org/projects/text-mining/) as well as code to parse and preprocess the resulting raw data. Such tools would extend our work here by making computational literature review available open-source, with no cost (except those imposed by JSTOR) and minimum coding skill requirements.

[11] View and contribute to our code at https://github.com/h2researchgroup/dictionary_methods and https://github.com/h2researchgroup/embeddings.

References

Allen, David M. 1974. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* 16, 125-127. (https://doi.org/10.1080/00401706.1974.10489157)

Alvesson, Mats, and André Spicer. 2019. Neo-institutional theory: A mid-life crisis? *Organization Studies*, 40 (2): 199-218.

Bail, Christopher A. 2014. "The Cultural Environment: Measuring Culture with Big Data." *Theory and Society* 43(3–4):465–82.

Bamman, David, Chris Dyer, and Noah A. Smith. 2014. Distributed representations of geographically situated language. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*: 828-834. Baltimore, MD, June 23-25.

Bartels, Larry M., J. Eric Oliver, and Wendy M. Rahn. 2016. Rise of the Trumpenvolk: Populism in the 2016 election. *ANNALS of the American Academy of Political and Social Science*, 667 (1): 189-206.

Benoit, Kenneth, Drew Conway, Benjamin E. Lauderdale, Michael Laver, and Slava Mikhaylov. 2016. Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review*, 110 (2): 278-295.

Bonikowski, Bart, and Laura K. Nelson. 2022. "From Ends to Means: The Promise of Computational Text Analysis for Theoretically Driven Sociological Research." *Sociological Methods & Research* 51(4):1469–83. doi: 10.1177/00491241221123088.

Braudel, Fernand. 1982 [1992]. *Civilization and Capitalism, 15th-18th Century, Vol. 2: The Wheels of Commerce*. (Translated by Siân Reynolds.) Berkeley: University of California Press.

Burt, Ronald S. 1983. *Corporate Profits and Co-optation: Networks of Market Constraints and Directorate Ties in the American Economy*. New York: Academic Press.

Carroll, Glenn R. 1985. Concentration and specialization: Dynamics of niche width in populations of organizations. *American Journal of Sociology*, 90: 1262-1283.

Davis, Gerald F. 2010. Do organizational theories progress? *Organizational Research Methods*, 13 (4): 690-709.

DiMaggio, Paul J., and Walter W. Powell. 1983. The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*, 48: 147-160.

Dodds, Peter Sheridan, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS One* 6 (12): e26752.

Garten, Justin, Joe Hoover, Kate M. Johnson, Reihane Boghrati, Carol Iskiwitch, and Morteza Dehghani. 2018. Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior Research Methods*, 50 (1): 344-361.

Geisser, Seymour. 1975. The predictive sample reuse method with applications. *Journal of the American Statistical* Association, 70: 320-328. (https://doi.org/10.1080/01621459.1975.10479865)

Graham, Jesse, Jonathan Haidt, and Brian A. Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96 (5): 1029-1046.

Greenacre, Michael. 2007. Hierarchical cluster analysis. *Correspondence Analysis in Practice*. London: Chapman and Hall/CRC.

Grimmer, Justin, and Brandon Stewart. 2013. Text as data: The promises and pitfalls of content analysis methods for political texts. *Political Analysis*, 21 (3): 267-297.

Hall, David, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. *Proceedings of the EMNLP* [*Empirical Methods in Natural Language Processing*].

Hannan, Michael T., and John Freeman. 1977. The population ecology of organizations. *American Journal of Sociology*, 82: 929-964.

Hannan, Michael T., and John Freeman. 1984. Structural inertia and organizational change. *American Sociological Review*, 49: 149-164.

Hannan, Michael T., and John Freeman. 1989. *Organizational Ecology*. Cambridge, MA: Harvard University Press.

Hannan, Michael T., László Pólos, and Glenn R. Carroll. 2007. *Logics of Organization Theory: Audiences, Codes, and Ecologies.* Princeton University Press.

Harris, Zellig S. 1954. Distributional structure. *Word*, 10 (2-3): 146-162.

Haveman, Heather A. 2022. *The Power of Organizations: A New Approach to Organizational Theory*. Princeton, NJ: Princeton University Press.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences*, 33 (2–3): 61-83.

Hsu, Greta, and Michael T. Hannan. 2005. Identities, genres, and organizational forms. Organization Science, 16 (5): 474-90.

JSTOR. 2018. JSTOR Title Lists. Retrieved October 11, 2018 (https://support.jstor.org/hc/en-us/articles/115007466248-JSTOR-Title-Lists).

Jurafsky, Daniel and James H. Martin. 2022. Chapter 19: Lexicons for sentiment, affect, and connotation. In *Speech and Language Processing (3rd Edition)*. Viewed October 1, 2022 (https://web.stanford.edu/~jurafsky/slp3/19.pdf).

Knorr-Cetina, Karen. 1999. *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge, MA: Harvard University Press.

Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84 (5): 905-949.

Kulkarni, Vivek, Rami Al-Rfou, Brian Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. *Proceedings of the 24th International Conference on World Wide Web*: 625-635. Geneva, Switzerland.

Laver, Michael, and John Garry. 2000. Estimating policy positions from political texts. *American Journal of Political Science*, 44 (3): 619-634.

Levy, Omer, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the ACL*, 3: 211-225.

Louwerse, Max M. 2004. Semantic variation in idiolect and sociolect: Corpus linguistic evidence from literary texts. *Computers and the Humanities*, 38 (2): 207-221.

Mabe, Michael, and Mayer Amin. 2001. Growth dynamics of scholarly and scientific journals. *Scientometrics*, 51 (1): 147-162.

Merton, Robert K. 1968. *Social Theory and Social Structure*. New York: Free Press.

Meyer, John W., and Brian Rowan. 1977. Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology*, 83: 340-363.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. Paper presented at the *International Conference on Learning Representations*, Scottsdale, AZ, May 2-4. (https://arxiv.org/abs/1301.3781)

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, Lake Tahoe, CA (December 2013): 3111-3119.

Miller, George A. 1985. Wordnet: A Dictionary Browser. in *Proceedings of the First Conference of the UW Centre for the New Oxford Dictionary*. Waterloo, Canada: University of Waterloo.

Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis* 16 (4): 372–403.

Moody, James, and Ryan Light. 2006. A view from above: The evolving sociological landscape. *American Sociologist*, 37 (2): 67-86.

Nectoux, François, and Fleur Thomése. 1999. Editorial. *European Journal of Social Quality*, 1: 3-11.

Nelson, Laura K., Derek Burk, Marcel Knudsen, and Leslie McCall. 2021. The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods and Research*, 50 (1): 202-237.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-Learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825-2830.

Pennebaker, James W., Cindy K. Chung, Molly Ireland, Amy Gonzalez, and Roger J. Booth. 2007. *The Development and Psychometric Properties of LIWC2007*. (www.LIWC.net)

Pfeffer, Jeffrey, and Gerald R. Salancik. 1978. *The External Control of Organizations: A Resource Dependence Perspective*. New York: Harper and Row.

Powell, Walter W., and Paul J. DiMaggio, eds. 1991. *The New Institutionalism in Organizational Analysis*. Chicago: University of Chicago Press.

Price, Derek J. de Solla. 1965. The scientific foundations of science policy. *Nature*, 206: 233-238.

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54 (1): 209-228.

Řehůřek, Radim and Petr Sojka. 2010. Software framework for topic modeling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*: 45-50.

Sapir, Edward. 1958. *Culture, Language and Personality*. Berkeley: University of California Press.

Schwartz, H. Andrew and Lyle H. Ungar. 2015. Data-driven content analysis of social media: A systematic overview of automated methods. *ANNALS of the American Academy of Political and Social Science,* 659 (1): 78-94.

Sivak, Elizaveta and Ivan Smirnov. 2019. Parents mention sons more often than daughters on social media. *Proceedings of the National Academy of Sciences*, 116 (6): 2039-2041.

Snefjella, Bryor, and Victor Kuperman. 2015. Concreteness and psychological distance in natural language use. *Psychological Science*, 26 (9): 1449-1460.

Stoltz, Dustin S., and Marshall A. Taylor. 2019. Concept mover's distance: Measuring concept engagement via word embeddings in texts. *Journal of Computational Social Science*, 2: 293-313.

Stone, Philip J., Dexter C. Dunphy, and Marshall S. Smith. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. Oxford, England: M.I.T. Press.

TensorFlow. 2018. Vector representations of words. (https://www.tensorflow.org/tutorials/representation/word2vec; retrieved 3 October, 2018)

Tetlock, Paul C. 2007. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62 (3): 1139-1168.

Towns, John, Timothy Cockerill, Maytal Dahan, Ian Foster, Kelly Gaither, Andrew Grimshaw, Victor Hazlewood, Scott Lathrop, Dave Lifka, Gregory D. Peterson, Ralph Roskies, J. Ray Scott, and Nancy Wilkins-Diehr. 2014. XSEDE: Accelerating scientific discovery. *Computing in Science & Engineering*, 16 (5): 62-74.

Turian, Joseph, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*: 384-394. Uppsala, Sweden: Association for Computational Linguistics.

Van Rossum, Guido, and Fred L. Drake. 2009. *Python 3 Reference Manual*. Paramount, CA: CreateSpace.

Vilhena, Daril A., Jacob G. Foster, Martin Rosvall, Jevin D. West, James Evans, and Carl T. Bergstrom. 2014. Finding cultural holes: How structure and culture diverge in networks of scholarly communication. *Sociological Science*, 1: 221-238.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, and Paul van Mulbregt. 2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17 (3): 261-172. (doi: 10.1038/s41592-019-0686-2)

Vogel, Rick. 2013. Visible colleges of management and organization studies: A bibliometric analysis of academic journals. *Organization Studies*, 33 (8): 1015-1043.

Whorf, Benjamin L. 1956. Science and linguistics. In John B. Carroll, ed., *Language, Thought and Reality: Selected Writings of Benjamin Lee Whorf*. Cambridge, MA: MIT Press.

Xing, Chao, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, Denver, CO, May 31 – June 5: 1006-1011.

Table 1: Example Expanded Dictionary: Organizational Ecology, 2004-2014

Term	Frequency	Cosine similarity
accountability	8782	0.145044
adaptation	10432	0.314533
adaptive	4192	0.304003
carrying_capacity	368	0.309869
competitive_intensity	436	0.348889
competitive_pressure	256	0.268243
competitive_pressures	397	0.268925
death_rate	269	0.215111
death_rates	385	0.247508
densities	529	0.250562
density	12211	0.322741
ecological	10359	0.302664
ecologists	344	0.296569
ecology	5623	0.292024
entry_exit	678	0.242189
evolution	13121	0.375792
evolutionary	6190	0.338242
expansion_contraction	74	0.244605
failure	17578	0.27059
fitness	1962	0.25413
founding	5897	0.331499
generalist	591	0.29947
generalists	444	0.292269
imprinting	538	0.274221
inertia	2435	0.342242
inertial	321	0.267593
legitimacy	16457	0.274898
liabilities_newness	55	0.284328

Table 1 (continued): Example Expanded Dictionary: Organizational Ecology, 2004-2014

Term	Frequency	Cosine similarity
liability_newness	170	0.324765
liability_smallness	62	0.246472
localized_competition	4146	0.314261
market_share	13548	0.241962
mortality	3565	0.262425
niche	1400	0.334246
niches	435	0.307522
organizational_change	3138	0.323099
organizational_form	1395	0.349514
population	80322	0.291708
populations	14106	0.260763
red_queen	128	0.274654
resistance_change	293	0.225986
resource_partitioning	312	0.36244
selection	29991	0.299911
senescence	63	0.240593
size_localized	36	0.301663
specialism	82	0.245359
specialist	1940	0.271816
specialization	4884	0.235251
survival	11936	0.307025
survive	3727	0.270919

Figure 1a:
Engagement in Sociology Journal Articles with Three Organizational Theories Over Time
(based on use of key concepts in journal articles)

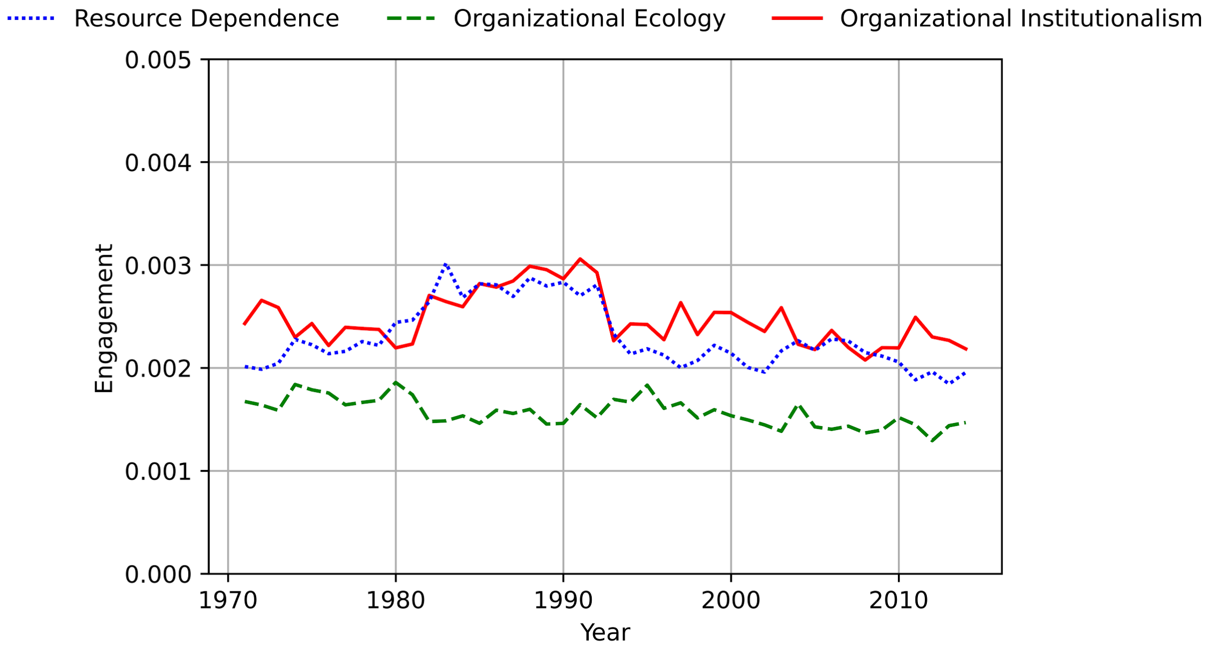


Figure 1b:
Engagement in Management Journal Articles with Three Organizational Theories Over Time
(based on use of key concepts in journal articles)

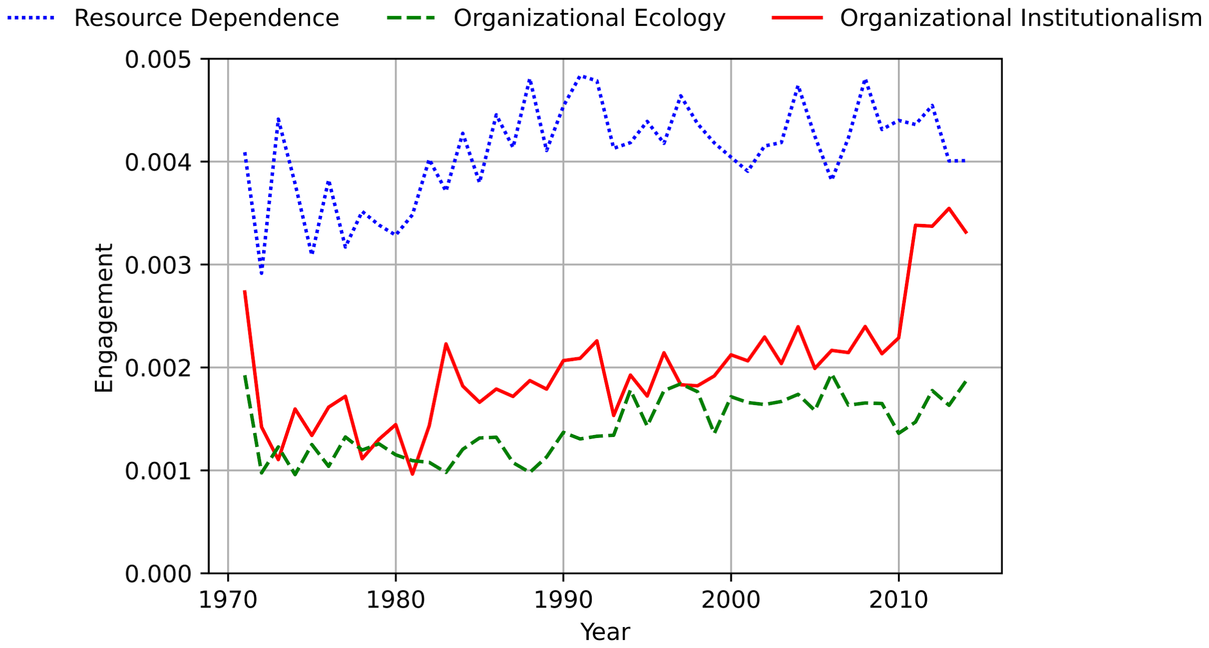


Figure 2: Engagement with Three Organizational Theories Over Time
(based on citations to foundational articles using Web of Science)

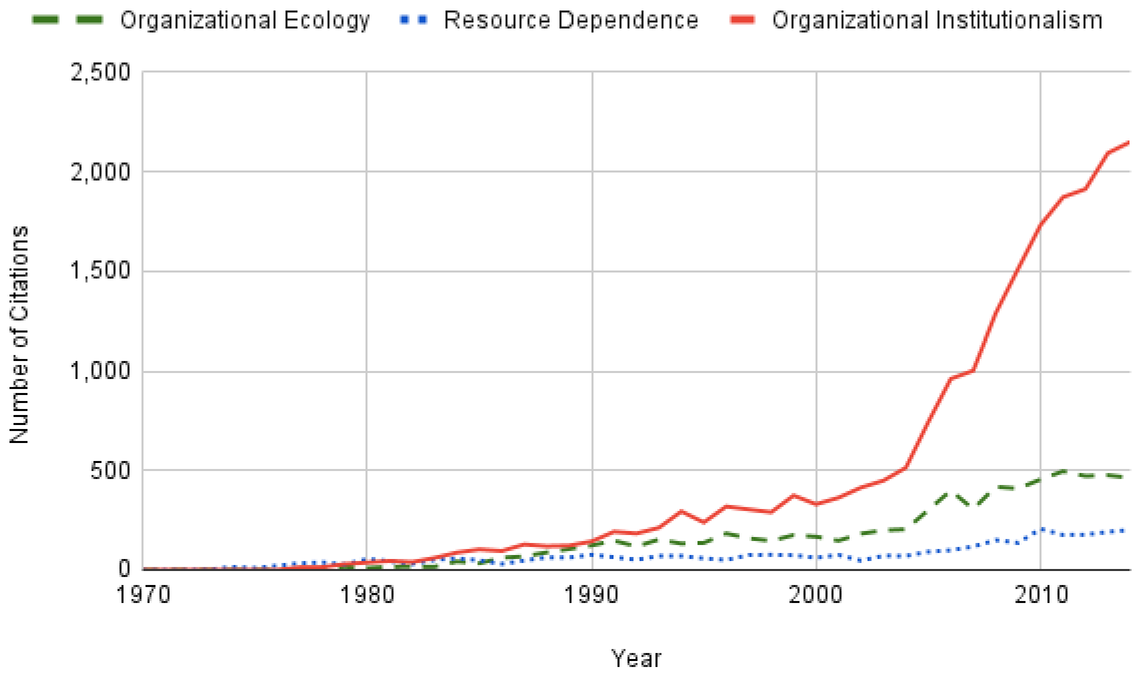


Figure 3a: Coherence of Three Organizational Theories Over Time (based on expanded dictionaries)

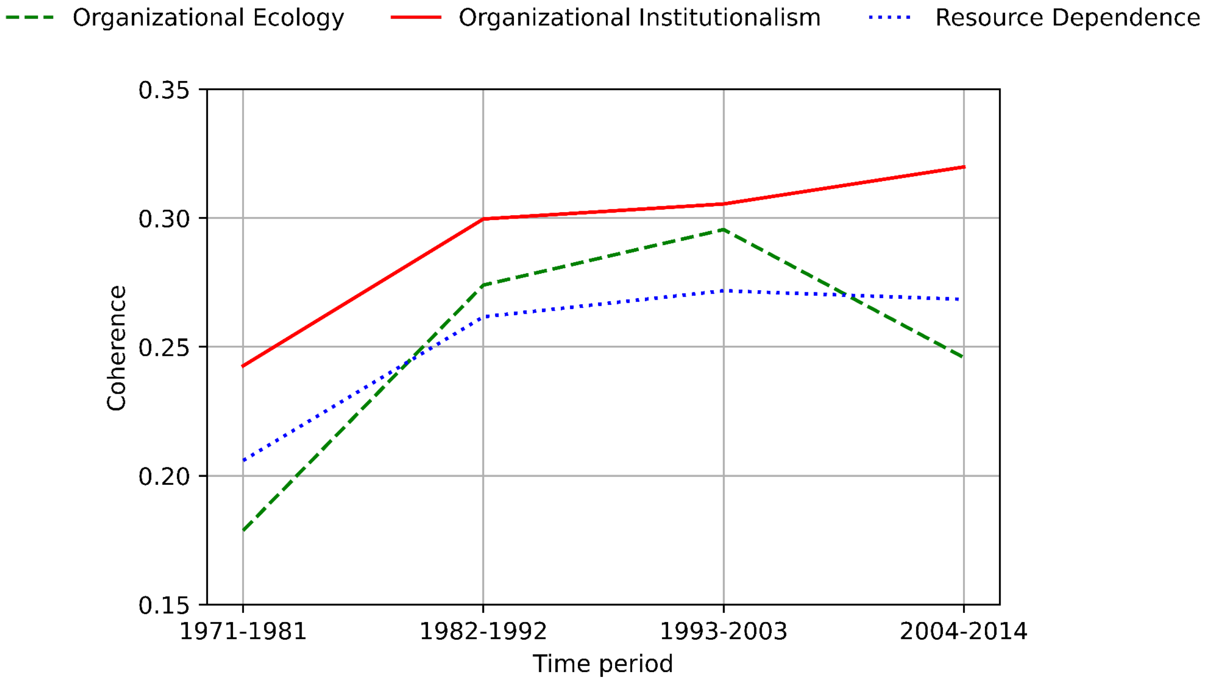


Figure 3b: Distinctiveness of Three Organizational Theories Over Time (based on expanded dictionaries)

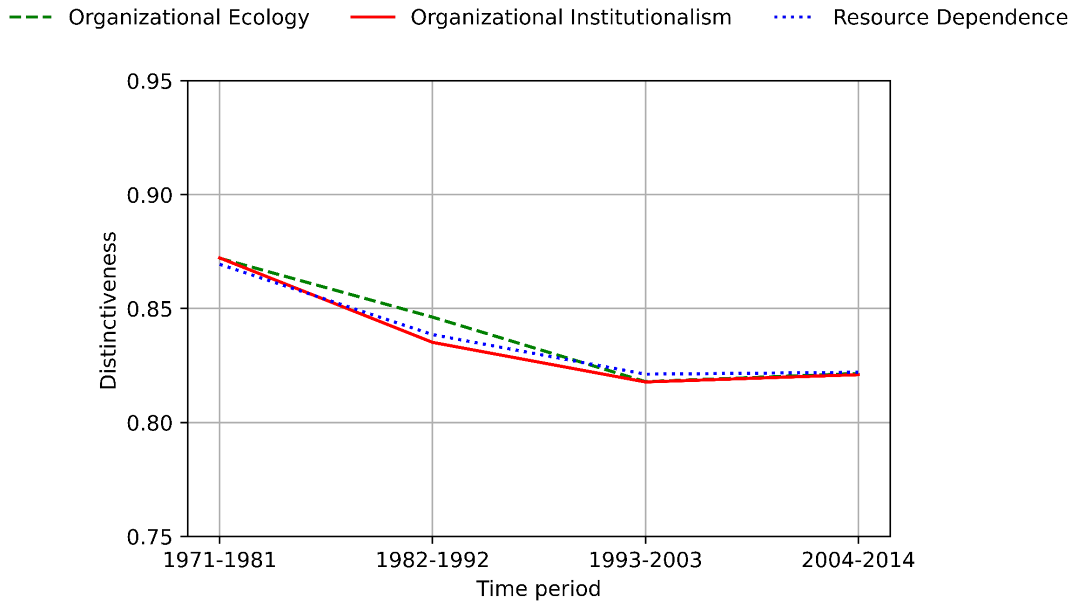
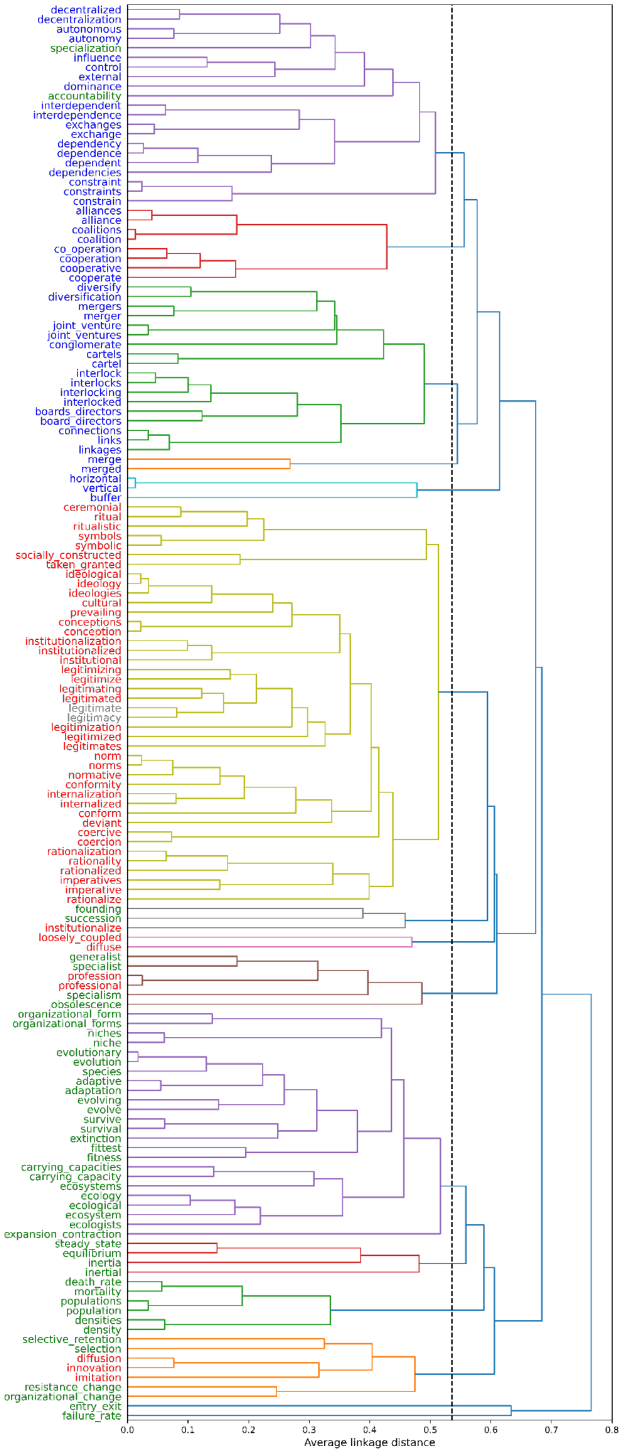
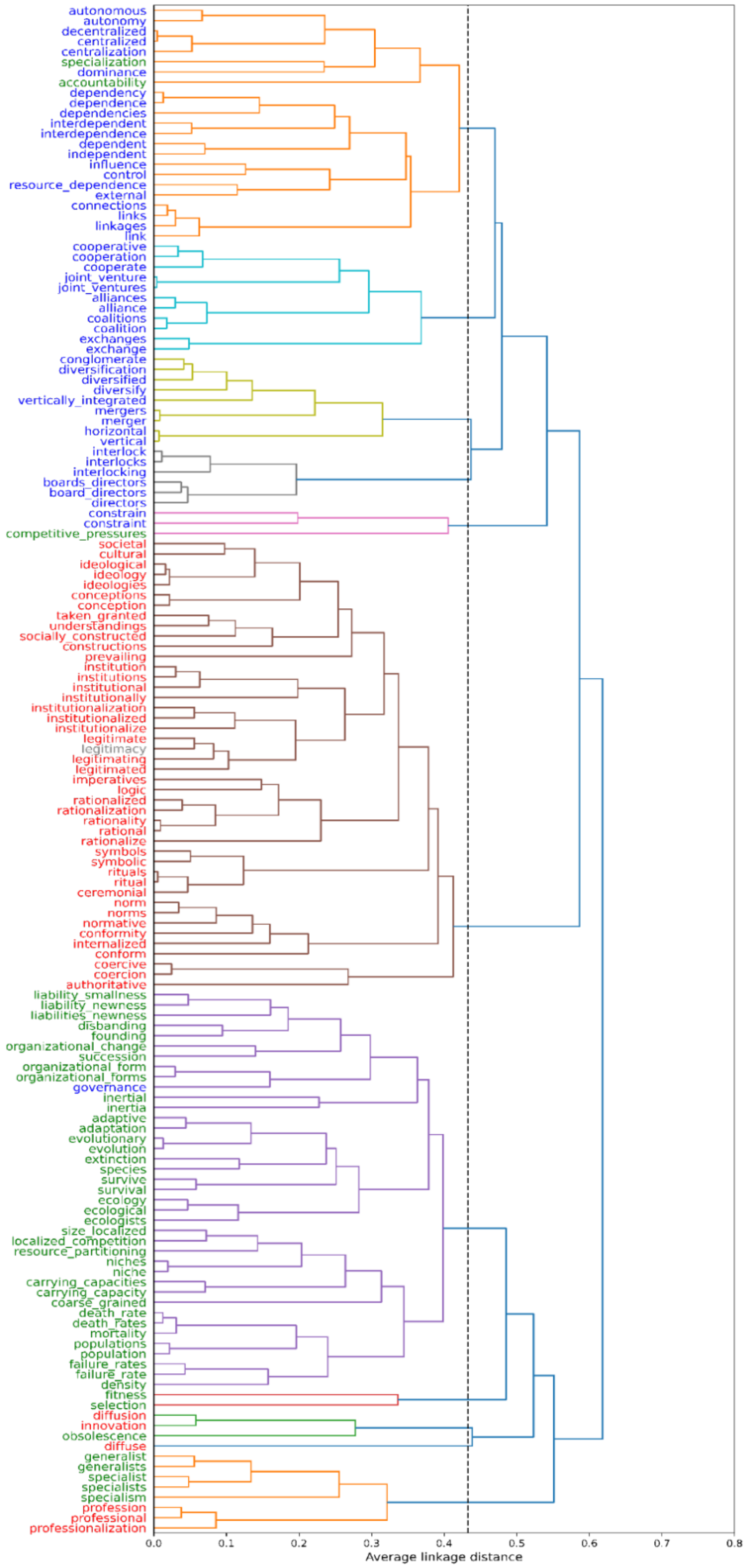


Figure 4a: Hierarchical Clustering of Expanded Dictionaries in 1971-1981



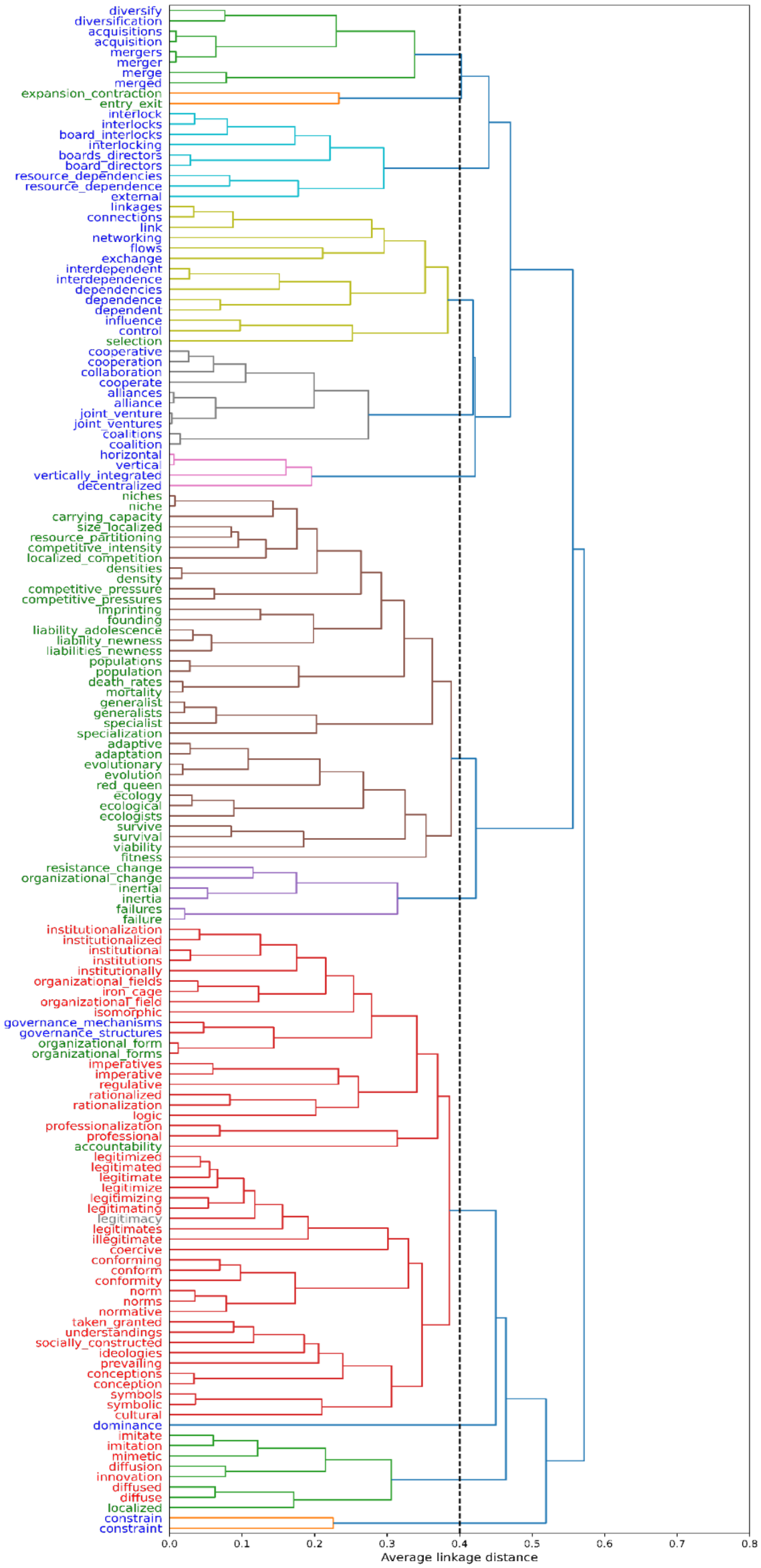
Note: Green indicates words denoting organizational ecology; blue indicates terms denoting resource dependence theory; red indicates terms denoting organizational institutionalism.

Figure 4b: Hierarchical Clustering of Expanded Dictionaries in 1982-1992



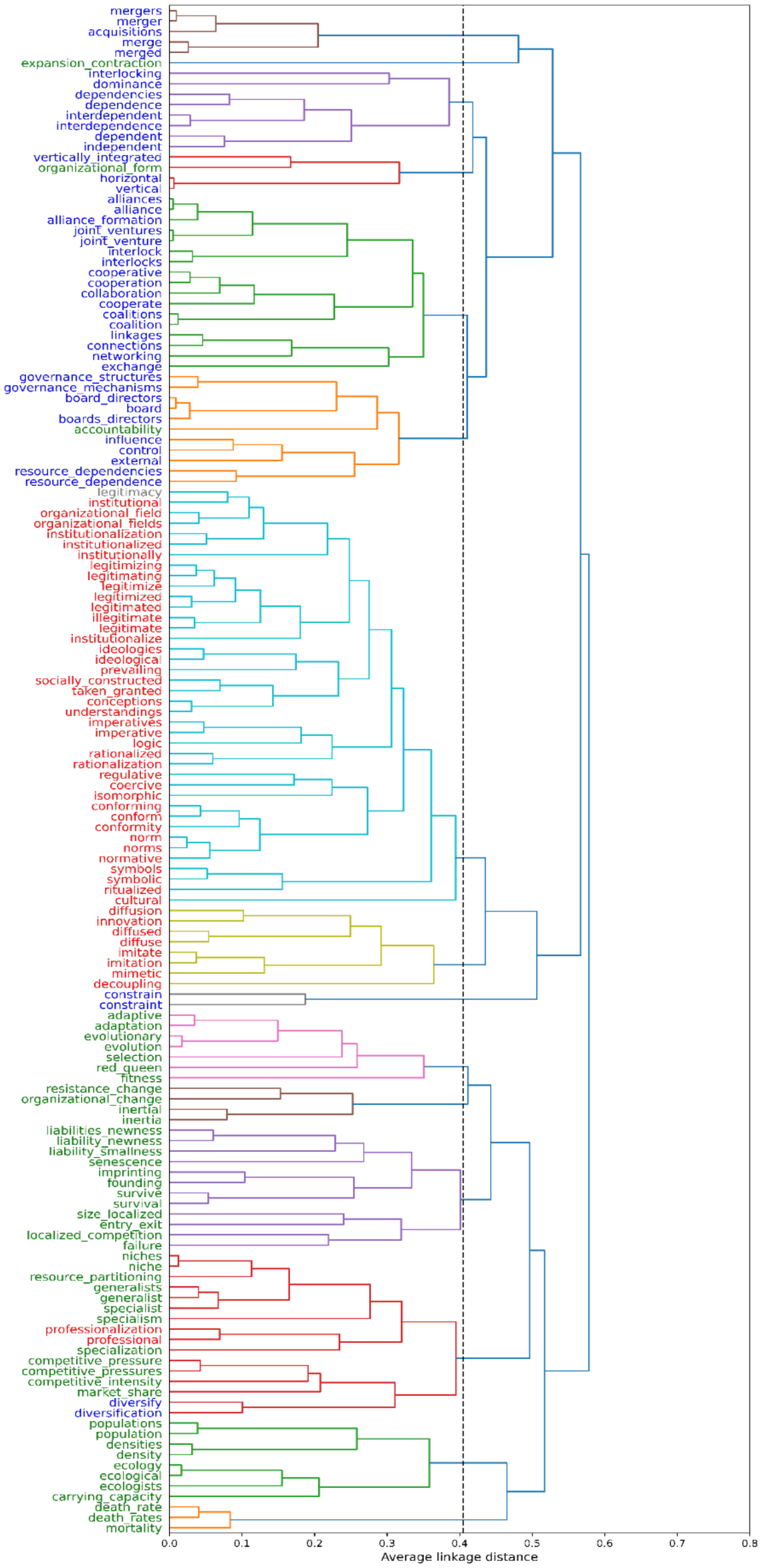
Note: Green indicates words denoting organizational ecology; blue indicates terms denoting resource dependence theory; red indicates terms denoting organizational institutionalism.

Figure 4c: Hierarchical Clustering of Expanded Dictionaries in 1993-2003



Note: Green indicates words denoting organizational ecology; blue indicates terms denoting resource dependence theory; red indicates terms denoting organizational institutionalism.

Figure 4d: Hierarchical Clustering of Expanded Dictionaries in 2004-2014



Note: Green indicates words denoting organizational ecology; blue indicates terms denoting resource dependence theory; red indicates terms denoting organizational institutionalism.

Technical Appendix

Details on Processing Raw Data from JSTOR

The raw data from JSTOR came in the form of 14 zip files, each about 1.5GB in size – 4.5GB when unzipped. For each article, the data include metadata (e.g., publication date, author name, journal title), counts of n-grams (single words and phrases of two or three words)[12], and a full-text file created through Optical Character Recognition or OCR (in .txt format). Although the OCR files are in raw format, JSTOR preprocessed the text data prior to counting n-grams. Specifically, JSTOR turned upper-case letters into lower-case letters, removed punctuation marks (including hyphens at the end of a line), replaced symbols such as hyphens and parentheses with white spaces, concatenated word sections separated by apostrophes, discarded common or stop words, and tokenized (but did not stem or lemmatize) the text using the Apache Lucene Standard Tokenizer according to standard unicode text segmentation practices.[13] The specific stop words removed were: "a", "an", "and", "are", "as", "at", "be", "but", "by", "for", "if", "in", "into", "is", "it", "no", "not", "of", "on", "or", "such", "that", "the", "their", "then", "there", "these", "they", "this", "to", "was", "will", and "with". Words were left in their original form rather than being stemmed (usually this means removal of word endings to leave a common base; e.g. "walking" and "walks" both reduce to "walk") or lemmatized (using morphology to index a word to an entry in a detailed dictionary; e.g., "studies" and "studying" are both linked to "study").

The JSTOR data contained many texts published in journals that were in languages other than English, many texts published in journals that were only secondarily related to sociology or management, and many texts that were not full-length articles. To eliminate extraneous texts, we filtered our corpus as follows. First, to limit the analysis to sociology and management, we identified the single most important subject for each journal using the JSTOR Complete Title History List (JSTOR 2018). We then forced each journal into a single primary subject area through rigorous manual checks informed by the second author's extensive knowledge of the field and, where necessary, inspection of journals and their contents. We assigned each journal to a primary subject area – sociology, management, or other. For example, *Administrative Science Quarterly* is listed under both sociology and management; the second author categorized this journal under management. Some journals' titles clearly indicated their primary subject; e.g., the primary subject for *Industrial and Labor Relations Review* is labor and employment relations, even though the journal is also listed under sociology and economics. A few titles required the second author to peruse journal contents or founding editorial statements. For instance, she categorized the *Journal of European Social Quality* under sociology rather than political science because its founding editors focused on "contemporary social issues in Europe," a combination of "the wellbeing of the individual person on the one side, and social cohesion, integration, and participation on the other" (Nectoux and Thomése 1999: 3), topics that are closer to sociology than political science. Finally, if a journal spanned subjects covered by this analysis and subjects outside this analysis (e.g., sociology and history)

[12] The analysis used bigrams and trigrams as well as unigrams because bigrams and trigrams capture complex ideas better than single words (Vilhena et al. 2014).

[13] For more details on preprocessing, see http://web.archive.org/web/20210506162859/https://www.jstor.org/dfr/about/technical-specifications (archived web page); https://solr.apache.org/guide/6_6/understanding-analyzers-tokenizers-and-filters.html; http://unicode.org/reports/tr29/#Word_Boundaries.

and if it had a significant footprint in either subject covered by this analysis, she categorized it as being under one of those subjects.

After defining journal subjects, we eliminated 160,202 articles in journals published in languages other than English,[14] and 82,948 additional articles from outside sociology and management. After these exclusions, which were based on the first author's experience and inspection of journals, we used metadata and (as a validity check) an unsupervised machine-learning algorithm, InferSent, to exclude 423 texts written in languages other than English. That left 155,553 texts published in 38 journals in management and 78 in sociology.

Details on the word2vec Algorithm

While two different model architectures have been proposed for word2vec, we used the "continuous skip-gram" model, which is the most accurate option for semantic comparisons (Mikolov, Chen, et al. 2013) – the focus of this paper – and is best suited for large data sets (TensorFlow 2018) like our academic articles corpus. In technical terms, continuous skip-gram seeks to classify a word given each other word in its context, allowing each word vector to have a separate impact on its neighbors. This is in contrast to the other principal word embedding architecture, the "continuous bag-of-words" model, which predicts target words based on their context, averaging all context word vectors, which treats the entire context as a single observation.

In our implementation of both word2vec, we took advantage of several extensions to the original continuous skip-gram model: noise reduction, undersampling of frequent words, and detection of common phrases in the corpus (Mikolov, Sutskever, et al. 2013). Common phrases, also called multi-word expressions, often possess unique meanings compared to their individual words; for instance, compare the meaning of the phrase "age dependent" to the individual terms "age" and "dependent". Our approach accepted phrases so long as their "collocation score" – the ratio of a given word pair's collocations divided by the product of each word's individual appearances – exceeded some threshold. Mathematically, this score is derived as follows:

$$\mathrm{score}(w_i, w_j) = \frac{\mathrm{count}(w_i w_j) - \delta}{\mathrm{count}(w_i) \times \mathrm{count}(w_j)}$$

where w_i and w_j are two words in the corpus, $w_i w_j$ is their collocation, and δ is a "discounting coefficient" that makes infrequent phrases less likely (Mikolov, Sutskever, et al. 2013: 6).

To be efficient, we used the gensim defaults for some parameters: 10,000 words per sample, 5 noise words (for noise reduction, which means ignoring words with frequency less than 5), an initial learning rate of 0.025, a discounting coefficient of 3, and a phrase detection threshold of 10. To improve the model's ability to capture semantic nuances without inducing unnecessary computational burdens, we expanded the defaults for a few other parameters: we used word context windows of size 10 to better capture syntactic relations (Mikolov, Chen, et al. 2013; Spirling and Rodriguez 2019; also optimal for doc2vec in Le and Mikolov 2014), 50 iterations through the data, and a hidden layer of size 300 (the most common choice). As suggested in Mikolov et al. (2013:4), we used a negative sampling exponent of 0.75 to subsample frequent words.

Because word-embedding models are stochastic – they are initialized with randomly chosen values – results can vary greatly across implementations, even when based on a single

[14] A tiny number of journals that publish articles in multiple languages, including English, were excluded (e.g., *Acta Historica Academiae Scientarium Hungaricae* and *Anabases*).

corpus (Tian et al. 2016; Hellrich and Hahn 2016b; Antoniak and Mimno 2018). This is especially likely when the corpus is small, in which case individual documents may have a large impact on the results. To generate robust estimates of word embeddings, we ran models repeatedly ("epochs" in NLP parlance). The gensim implementation of word2vec in Python allows researchers to train models iteratively. The first epoch is initialized with random values for parameters. The output for the first epoch is used as input to the second epoch, the output for the second epoch is used as input to the third, and so on. We continued this cycle for 50 epochs. There is no clear standard for judging when a custom-trained word2vec model is robust, so we followed previous research (e.g., Kulkarni, Al-Rfhou, Perozzi, and Skiena 2015; Hellrich and Hahn 2016a) and used 0.990 as a threshold. For all time periods, the word2vec models reached that threshold before 50 epochs.

References (only those that are not in the main paper)

Antoniak, Maria, and David Minmo. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6: 107-119.

Hellrich, Johannes, and Udo Hahn. 2016. An assessment of experimental protocols for tracing changes in word semantics relative to accuracy and reliability. *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*: 111-117.

Hellrich, Johannes, and Udo Hahn. 2016a. An assessment of experimental protocols for tracing changes in word semantics relative to accuracy and reliability. *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*: 111-117.

Hellrich, Johannes, and Udo Hahn. 2016b. Bad company—Neighborhoods in neural embedding spaces considered harmful. *Proceedings of the 26th Annual Conference on Computational Linguistics*: 2785-2796.

Le, Quoc, and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China, June 22-24. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 32 (2): 1188-1196. (https://arxiv.org/abs/1405.4053)

Spirling, Arthur, and Pedro L. Rodriguez. 2019. Word embeddings: What works, What doesn't, and how to tell the difference for applied research. Working paper, NYU Department of politics.

Tian, Yingtao, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2016. On the convergent properties of word embedding methods. (https://arxiv.org/abs/1605.03956)

Figure A1a: Coherence of Three Organizational Theories Over Time (based on seed dictionaries)

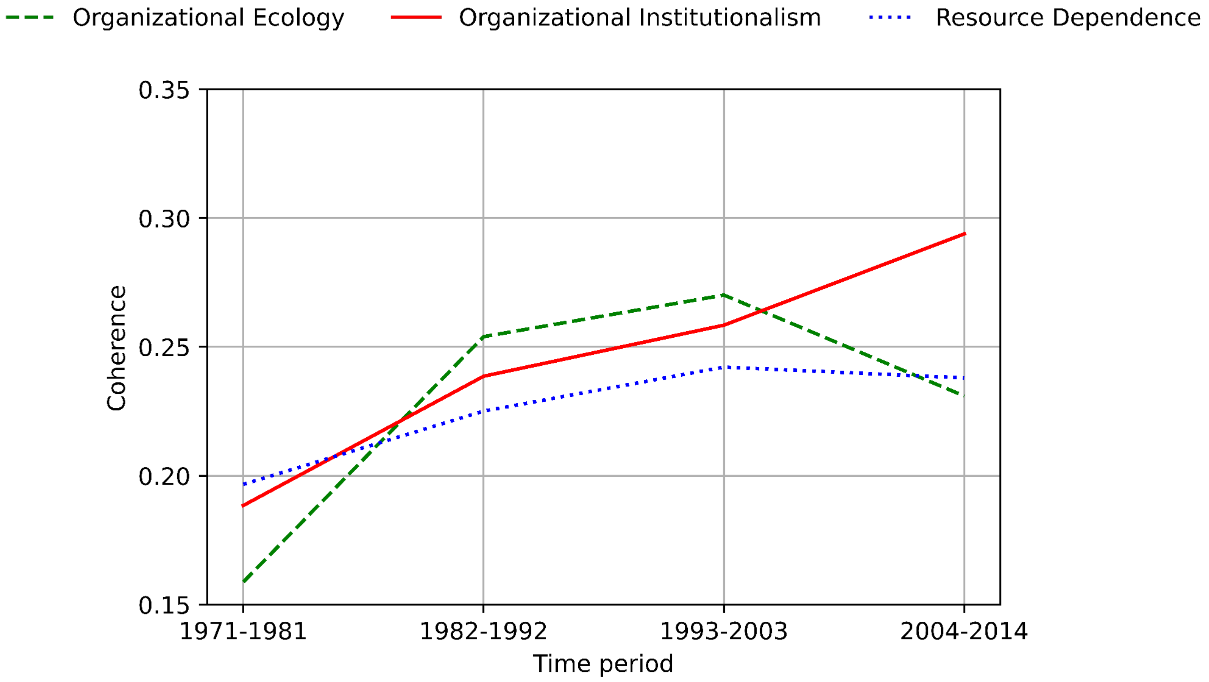


Figure A1b: Distinctiveness of Three Organizational Theories Over Time (based on seed dictionaries)

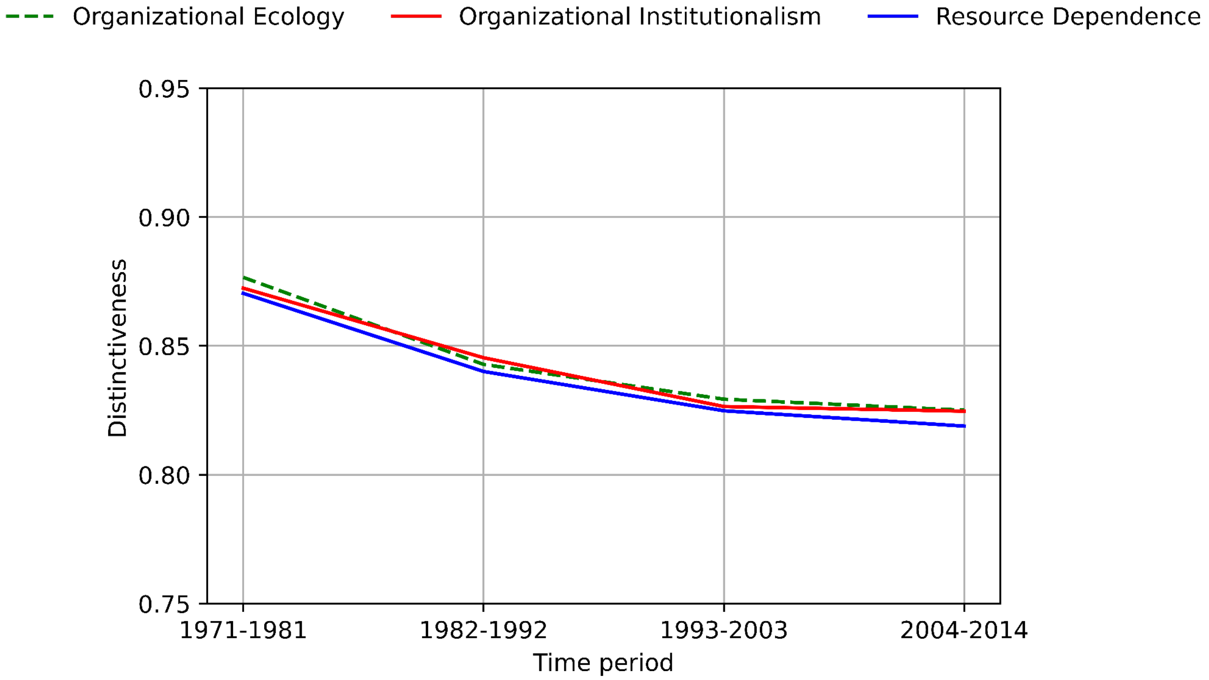


Figure A2: Engagement with Three Organizational Theories Over Time (based on citations to 30 distinctive articles per theory using Web of Science)

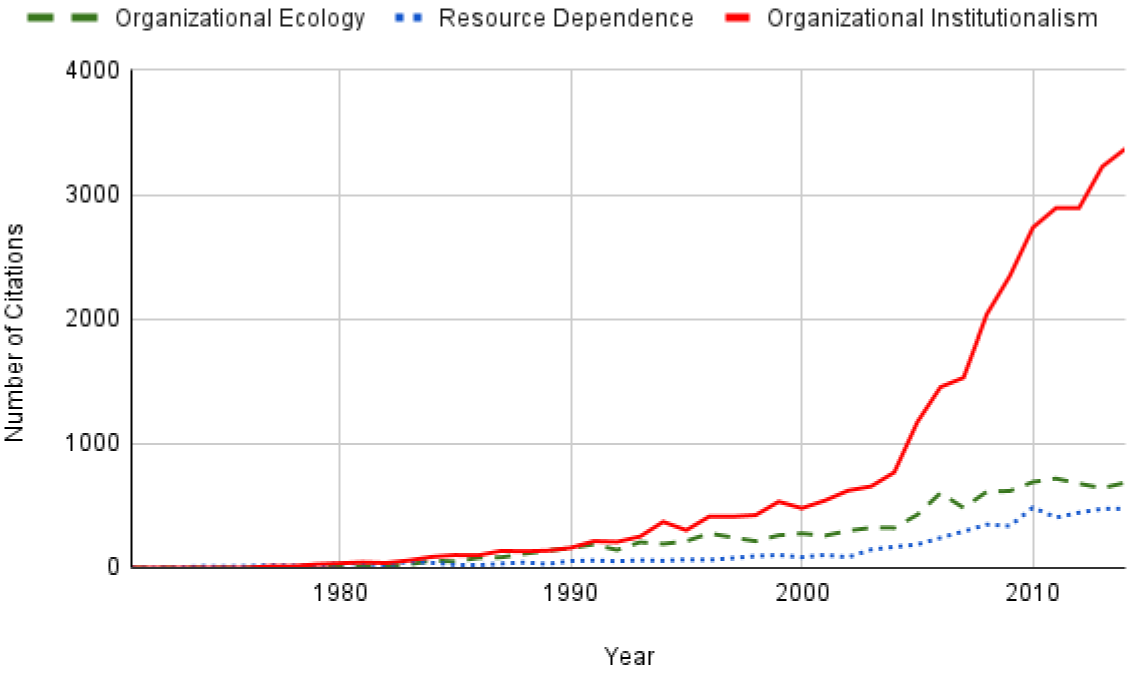


Figure A3a:
Coherence of Three Organizational Theories Over Time (based on seed dictionaries)

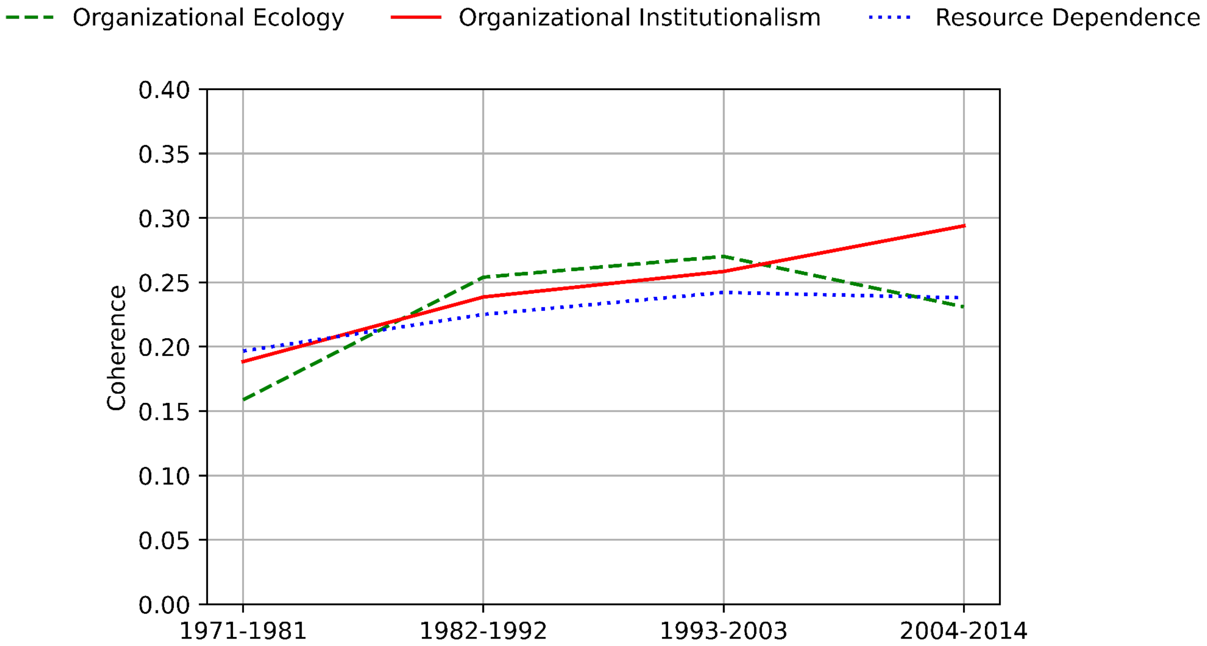
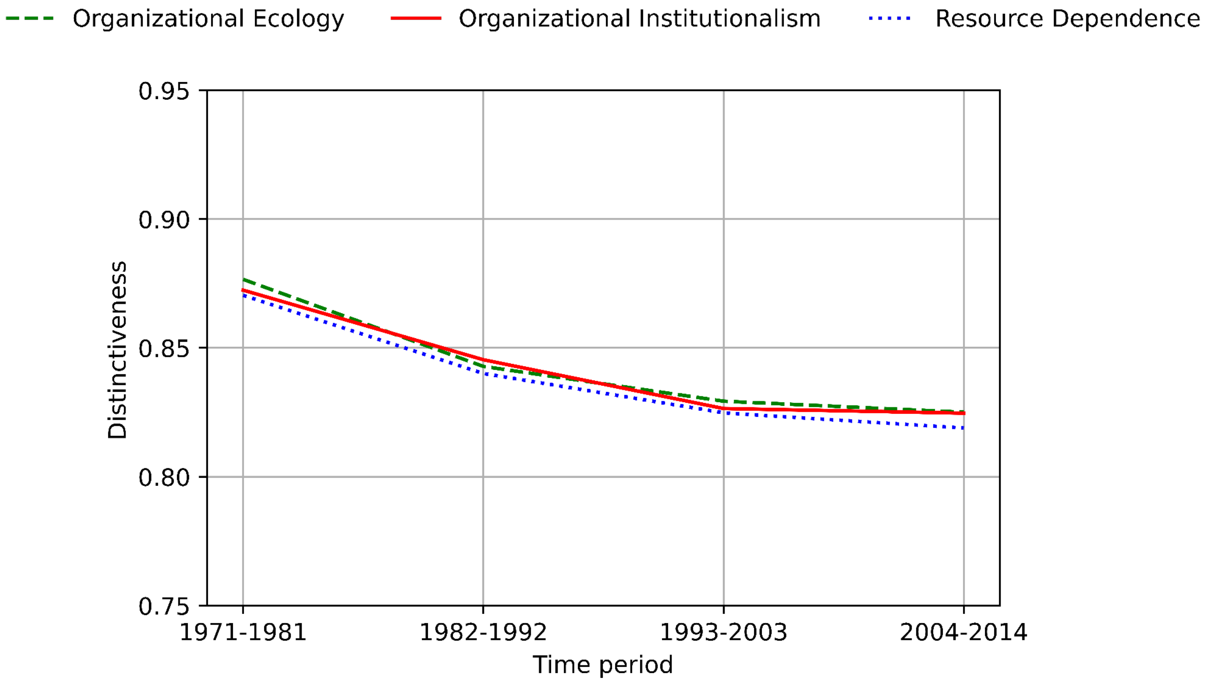


Figure A3b: Distinctiveness of Three Organizational Theories Over Time (based on seed dictionaries)


Organizational Ecology
Organizational Institutionalism
Resource Dependence
Distinctiveness
Time period