THE UNIVERSITY OF CHICAGO

Examining Semantic Shifts of Sociological Papers from 1940 to 2020 Using Text Embeddings

By

Tianle Ye

May 2025

A paper submitted in partial fulfillment of the requirements for
the Master of Arts degree in the Master of Arts in
Computational Social Science

Faculty Advisor: Ali Sanaei

Preceptor: David A. Peterson

Abstract

Concerns have been raised regarding the potential fragmentation and weakening historical continuity within sociology during the twentieth and twenty-first centuries. This study investigates the evolution of sociological discourse by using text embedding models. By turning paper documents into high-dimensional embeddings, this research analyzes eight decades of papers (1940-2020) from the *American Sociological Review* (ASR) and the *American Journal of Sociology* (AJS). Findings indicate that while semantic meanings of sociology gradually and steadily shift away from its past, the studies in the twenty-first century have only tenuous semantic relationships with those in the 1940s and 1950s. Moreover, the development of sociology appears to follow the shape of multiple ongoing subprocesses rather than a series of clearly defined distinct periods. This research contributes to understanding the semantic shift of publications in social sciences.

Keywords: sociology of sociology; sociology of knowledge

1 Introduction

How did different generations of sociology evolve during the twentieth and twenty-first centuries? How do they relate to their predecessors? Has sociology weakened its connection with its historical past? On the surface, the development of sociology during past decades appears to be characterized by a succession of distinct periods. However, the chronological semantic shifts of sociological work might be more intricate and nuanced. As some scholars have argued, sociology has become increasingly fragmented and disintegrated, raising concerns that the discipline may be losing its theoretical core and intellectual cohesion (e.g., Turner, 1989; Stinchcombe, 1994). This indicates that sociology might be better characterized by multiple overlapping subprocesses unfolding within its subfields. It also indicates that the lines between different generations of sociology might have been blurred in sociology.

By using text embedding methods, this research attempts to respond to these ideas on sociology. Based on eight decades of papers published in the *American Sociological Review* (ASR) and the *American Journal of Sociology* (AJS) from 1940 to 2020, the current research finds that while semantic meanings of sociology gradually and steadily shift away from its past, the studies in the twenty-first century have only tenuous semantic relationships with those in the 1940s and 1950s. Moreover, the development of sociology appears to follow the shape of multiple ongoing subprocesses rather than a series of clearly defined distinct periods. This thesis will begin with a literature review on the historical development

of sociology that contextualizes my research questions, followed by an overview of text embedding applications in the social sciences that contextualizes my use of document-level embeddings. Then I will introduce my two main datasets, papers from ASR and AJS, and explain my application of document-level embeddings. Next, I will show the results I generated from different years' text embeddings and interpret their meanings in relation to the research questions. The thesis will conclude by outlining the limitations of the current research and proposing potential directions for future inquiry.

The reason I do this thesis is that the questions and concerns I propose here have their normative meaning. If sociology is increasingly detached from its past, it might lose its disciplinary integrity and future direction. If sociology is composed of multiple subprocesses and specialized discussions whose influence tends to be short-lived, this may suggest that the field is noncumulative and fragmented. Losing disciplinary core and being fragmented are not good signs for an academic discipline. However, as multiple scholars have articulated, sociology is increasingly faced with the problem of disintegration and fragmentation (e.g., Turner, 1989; Stinchcombe, 1994). This thesis is part of the effort to unveil what has been going on in sociology empirically, thereby giving foundations to such criticisms of the field.

2 Literature Review

2.1 A short history of sociology: different periods and forgotten past

The history of sociology since 1940 can be roughly divided into a number of periods. Sociology from the 1940s to 1960s was increasingly influenced and dominated by Parsons' functionalism theory, a grand macro-level social theory that focuses on social order and functions of social institutions. Impacted by various external social events in the late 1960s, the order-centered Parsonian theory increasingly faced criticisms from scholars from young generations and gradually lost its validity and social relevance (Joas and Knöbl, 2009). In substitution of the dominance of Parsonion theory, a number of competing theoretical perspectives and paradigms emerged during the late 1960s and early 1970s. Notable examples include neo-functionalism, symbolic interactionism, conflict theory, exchange theory, structuralism theory, and phenomenological theory (Turner, 1989). At the same time, influenced by Merton's middle-range theory, the same period also witnessed the flourishing of empirical work that focused on specific social fields and social institutions. Entering the 1980s and 1990s, these empirical studies gradually turned to the topics of identity and culture under the influence of feminism and the "cultural turn" (Susen and Turner, 2011). The same period also saw the emergence of a number of grand social theories that attempted to synthesize the findings in the field such as Habermas's theory of communicative action (2007), Giddens' structuration theory (1986), and Bourdieu's theory of social and cultural

capital (2002). Such productions of grand social theories gradually came to a halt in the twenty-first century. What increasingly replaced theory production was the growth of empirical work, particularly in North America. The topics of sociological empirical research consistently expanded in the new century, leading some scholars to believe that sociology was in a stage of disintegration and fragmentation (e.g., Turner, 1989; Stinchcombe, 1994).

This description of sociology's past suggests that the development of sociology from the 1940s until now can be roughly divided into a number of distinct periods. Indeed this is the mode in which some scholars describe the history of sociology (e.g., Turner, 1989). Such a mode of periodization has led some researchers to conclude that social sciences such as sociology have clear cycles of development. For example, Abbott's fractal model of the development of academic disciplines suggests that the history of sociology can be regarded as a series of fracturing processes, each lasting for two to three decades (Abbott, 2001). However, such periodization might merely serve as a historical method rather than capturing the actual trajectory of sociology. It should be noted that when scholars talk about different periods of sociology, they tend to refer to a number of schools of thought or theoretical perspectives as their defining attributes. To be clear, they typically don't discuss all research published within a particular period collectively; rather, they tend to focus only on representative works that are highly influential and capable of defining an era. In reality, researchers in a given era might be doing something entirely different from those works that define that generation of researchers. The scope and the significance of such researchers' academic productions should not be underestimated. Therefore, it is possible that the periodization of sociology might be an artifact if we take the research published as a whole.

In fact, such a method of periodization may better be applied to some specific branches and areas of research in sociology. Indeed, the core of Abbott's model (2001) of scholarship periodization, the fractal model, may be best to describe the dynamics of some subprocesses in sociology. For Abbott (2001), an academic field is supposed to be organized by a number of binaries such as qualitative and quantitative, macro and micro. As these oppositions debate and argue with each other, a victory paradigm would gradually emerge from one side substituting the old paradigm. Having won, this paradigm would give birth to new binaries and this whole fractal process would repeat itself. However, it should be noticed that such a dominant paradigm or consensus has long disappeared from sociology. Sociology as a general subject has been increasingly regarded as fragmented, disorganized, and nonconsensual (e.g., Turner, 2001). Nevertheless, this pattern of research cycle might be identified within some subfields in sociology. Many specific groups of research such as the discussion of institutionalism (e.g., Meyer and Rowan, 1977; DiMaggio and Powell, 1983) start with a debate that lasts for several years and ends with a more or less clear consensus. Such

discussion may not give rise to another round of debate concerning a topic that pertains to it. New discussions on different topics may take the place of earlier academic debates, thus carrying forward the historical development of sociology. Therefore, the development of sociology may be more accurately described as a set of overlapping and evolving subprocesses that emerge and decline over time, rather than as a neatly segmented periodization of the discipline as a whole.

As sociology gradually developed and differentiated, questions arose as to whether the discipline tends to lose sight of its past traditions. To be sure, there are a number of influential theoretical syntheses in the 1980s and 1990s that attempted to revitalize the sociological tradition and apply them for timely theoretical articulation. Such notable examples include Jeffery Alexander's neo-functionalism (2014) and Giddens' structuration theory (1986). However, if posed with this question, many scholars today would agree that current sociology is increasingly detached from its past. For example, Turner (2001) argues that in sociology "the defense of specific disciplinary traditions and objectives has been strangely neglected" (pp. 5). This neglect even prompted Turner to establish the *Journal of Classical Sociology* as a way to safeguard the discipline's tradition. Such separation from the discipline's past has also been found in classroom settings. In a study of 35 textbooks in sociology, Keith and Ender (2004) found that those textbooks in the 1940s and 1990s only shared less than three percent of all concepts. I doubt such a number would occur in political science or economics. Such departures from past traditions might undermine the integrity of sociology as a discipline and inhibit researchers' ability to question, hindering the overall development of the field (Szelenyi, 2015).

Based on the discussion above, three general questions can be asked for my current research. First, does the development of sociology follow a period-and-transition pattern or a gradual and steady pattern? Second, is it so that rather than having a clear period-after-period pattern for sociology in general, sociology is more characterized by multiple specific academic discussions that occur and diminish? Third, has current sociology gradually lost its connection with the past? Three hypotheses can be generated from these three questions:

H1: *The semantic evolution of sociological research follows a gradual and continuous trajectory over time.*

H2: *Sociology consists of the rise and decline of specific thematic or theoretical discussions whose timespans are limited.*

H3: *Contemporary sociological research has been semantically separated from the earlier writings in the past.*

In the following research, I will use text embedding models to turn these texts into embeddings that represent each paper's meaning. To understand this method, it is first important to understand what text embedding is and how it has been applied in social

sciences.

2.2 Applications of text embeddings in social sciences

My use of text embeddings builds on earlier applications of this technique in social scientific research. A text embedding is a numeric vector that represents the meaning of a text based on the semantic meaning of the embedding model learned from a given corpora. By turning unstructured textual data into structured vectors, text embeddings enable researchers to identify the positions of texts in the high-dimensional meaning space and explore how different texts relate to each other in that space. It also enables researchers to compare the meanings of different texts mathematically with methods such as the computing of cosine similarity. Traditional text embedding method uses word embedding models such as Word2Vec and GloVe to represent the meaning of each word in a text by a vector. These word embedding models allow researchers to examine different qualities of the concepts such as conceptual breadth, similarity, meaning, and positions within cultural and knowledge continua (Aceves and Evans, 2024). With the rise of transformer-based language models such as BERT (Bidirectional Encoder Representations from Transformers), text embeddings can now capture semantic information at the phrase and sentence levels. Beyond the scope of sentences, current state-of-the-art embedding models that are based on Large Language Models (LLMs) can even turn the semantics of an entire document into a single vector representation. These sentence and document-level embeddings retain the contextual meaning of texts and therefore can be used to compare and assess how similar two pieces of texts are in terms of their general meanings.

Currently, word embedding models have been widely employed in the field of social sciences. This is exemplified by the work of Kozlowski et al. (2019) in which the authors invented a novel method for getting cultural dimensions. The core idea of their approach is by adding or subtracting two-word vectors, one can get the meaning of a two-word phrase or a meaning dimension if the two words subtracted are binary. For example, if one wants to get the meaning of black woman in the vector space, a mere summing of the word black and the word woman is enough to create the meaning space for black woman. To give another instance, if one aims to get the dimension of gender, one can find relevant word pairs such as man-woman and subtract the vector of man by women. Having this dimension, one can then check a word's meaning in relation to a meaning dimension by calculating their cosine similarity. Researchers have also applied the word embedding model to study a variety of cultural elements such as morality (Arseniev-Koehler et al., 2022), status (Peng et al., 2021), intersectionality (Nelson, 2021), and ideology (Taylor and Stoltz, 2021). Beyond word embeddings, transformer-based models that produce sentence or document-level embeddings are also increasingly used in social sciences. Vicinanza et

al. (2023) utilized fine-tuning BERT model to compute contextual novelty for sentences in corpora of politics, law, and business. Bestvater and Monroe (2023) built a classifier based on BERT sentence embeddings to detect political stances of short texts from tweet if they approve or oppose feminism movements. Licht (2023) used multilingual sentence embeddings to categorize ideological positions and topics of party manifestos. Lin (2025) applied BERT-based cross-encoders to estimate the semantic similarity of political texts.

The current study turns different papers into embeddings with a frontier embedding model, Google Gemini embedding model experiment 03-07, which produces text embeddings of more than 3,000 dimensions. With such high dimensions, the text embeddings are supposed to retain rich meanings for the original texts. Moreover, the structured data format of vector representation enables researchers to compare the different texts' meanings quantitatively. By turning all relevant academic works into embeddings, this work also avoids the limitation of analyzing only a few representative texts from each era and therefore can represent sociological work comprehensively. These are the reasons I use text embeddings to study the questions I proposed above.

3 Data and Methods

3.1 Data

For the current research, I use two top journals in the field of sociology in America: the *American Journal of Sociology* (AJS) and the *American Sociological Review* (ASR). I choose these two journals for the following reasons. First, these two journals represent the best research and the frontier in the field of sociology. Second, compared with other journals in the field, these two journals have a relatively long history that allows for research on the historical development of the field. While AJS can be dated back to as early as the 1890s, ASR can be traced back to the 1930s. This timeframe allows for research over eight decades of sociological publications. Third, different from many sociological journals that only focus on a specific subfield (e.g., gender, race, family), these two journals encompass a variety of important sociological topics that are representative across the field. One of the key problems of new sociological fields is that their field names may only be an empty label void of substantive content (Starr, 1983). The selection of these two top journals serves as a gatekeeper, assuring that the topic labels entered into the field are nontrivial and have substantive content.

To fetch the documents in the two journals, I first took advantage of Crossref's metadata of academic journals and successfully downloaded all DOIs from the two journals. In total, I fetched 13,168 DOIs from ASR and 27,592 DOIs from AJS. To get the paper PDFs from AJS, I first contacted the University of Chicago Press and received data mining permission, then I

used selenium to automatically collect all of its online publications with their corresponding DOIs. For ASR, I applied a Python scrapping library oafuncs to automatically scrap most of its papers and then manually download those that are not downloaded.

I then separated papers from the rest of the other texts as I found that the publication of book reviews and other types of texts varies from year to year in the two journals. For instance, during its early decades, the American Sociological Review regularly published a number of book reviews, while in recent years, such reviews have vanished from the journal. However, the publication of papers remains consistent across time for both journals. Moreover, the writing style of book reviews and other texts has more variations and is distinct from the style of the papers. Including book reviews might therefore involve unnecessary writing style differentiation for a given corpora. Consequently, I only retained the journal papers and discarded the rest of the texts. I select papers from the two journals by looking at their titles. If the titles include italicized book titles or expressions that apparently indicate a non-paper item (e.g., erratum, in memoriam of), I would then discard them. I then selected the papers published between the years of 1940 and 2020. Together, I retained 4,181 papers from ASR and 3,620 papers from AJS.

Next, to get texts from the PDFs, I used an open-source PDF converter called MinerU which allows for converting PDFs into JSON files. In these JSON files, each separate dictionary restores a section of text that is segmented by the OCR, the text type (text or table/figure) and text level (title or normal text) of the text segmentation, and its page number. Such a structure of the results from PDF recognition enables me to extract redundant patterns of texts with higher accuracy. I then identified the patterns of redundancy in the JSON files and removed them. Two important instances of such redundancy are the front page of some ASR papers, which consists of only metadata, and the appendixes and references at the end of each paper. I removed them either by regex or by the mark of the section (e.g., “REFERENCES”). Removing these excessive parts, I then gathered the whole text from the JSON files for the papers. I then did basic text cleanings on these raw texts, removing elements such as URLs, email addresses, special characters, and redundant punctuations. I didn’t use lowercase or lemmatization to retain the basic semantic structure of the texts. The cleaned texts still have a small portion of misspellings and concatenated words (e.g., the word). I tried to solve these issues by using existing libraries in Python or large language models. While the former only complicates the issue by introducing more noise, the latter is too time and resource-consuming. Considering that these misspellings and word concatenations only consist of a tiny portion of my texts and that the Gemini embedding model is designed to handle such issues, I opted not to make more corrections to the cleaned texts.

3.2 Methods

To get the embeddings for the papers, the first step is to chunk them into different sections that fit the token limit of the embedding models. The upper limit of token number of my Gemini embedding model is 8,000. I used Langchain's RecursiveCharacterTextSplitter to chunk each paper into sections no more than 7,950 tokens to leave room for system tokens and set an overlap ratio of 0.05 for each chunk to better preserve its semantic coherence. I used this specific splitter because RecursiveCharacterTextSplitter can help retain the semantic structure for different chunks. After I got the chunks for each paper, I used one of the frontier embedding models, gemini-embedding-exp-03-07 from Google, to get the embedding for each chunk with a Google API. The gemini-embedding-exp-03-07 model can generate an embedding of 3,000 dimensions. As such, the embeddings can authentically capture the semantic meaning of the original texts. After getting all the embeddings, I then averaged them for each paper to get embedding vectors that represent the meanings of different papers.

To answer my research questions, I first averaged all the text embeddings by different years to get each year's centroid of embeddings. These centroids of embeddings represent the general semantic meaning of all published papers of a given year. Then I used Uniform Manifold Approximation and Projection (UMAP) dimension reduction to visualize each year's embedding centroid. As UMAP has a good balance preserving both global and local structures when visualizing the embeddings, it can give me first a sense of the general pattern of semantic distribution of different years' publications. I set n_neighbors as 15 to better balance local and global structure and min_dist as 0.5 to give moderate separation without overcrowding points. I also set a random state in order to replicate the result. I then applied the hierarchical clustering strategy and calculated the year-by-year cosine similarity score for the embedding centroids. I used linkage and Ward method because it keeps each merge between years as tight as possible and therefore produces clear and compact clusters. The resulting cosine-similarity matrices were visualized as paired heat maps for easy comparison. These two methods are used to discover the semantic similarity pattern of all years' published work. If the semantics of different years are clustered in the same group or are colored by the colors that represent high cosine similarity, those years are supposed to be semantically similar. I also calculated the Euclidean and cosine distances for embedding centroids between the starting year of 1940 and every subsequent year as well as between each pair of consecutive years. Whereas Euclidean distance focuses on the magnitude of change in meanings, cosine distance gauges how closely the semantics of two years align or diverge. Combined together, these two metrics capture both the degree and the direction of semantic change across the corpus over time. In addition, I also used UMAP to visualize all papers' embeddings by different decades to see how as a whole the semantics

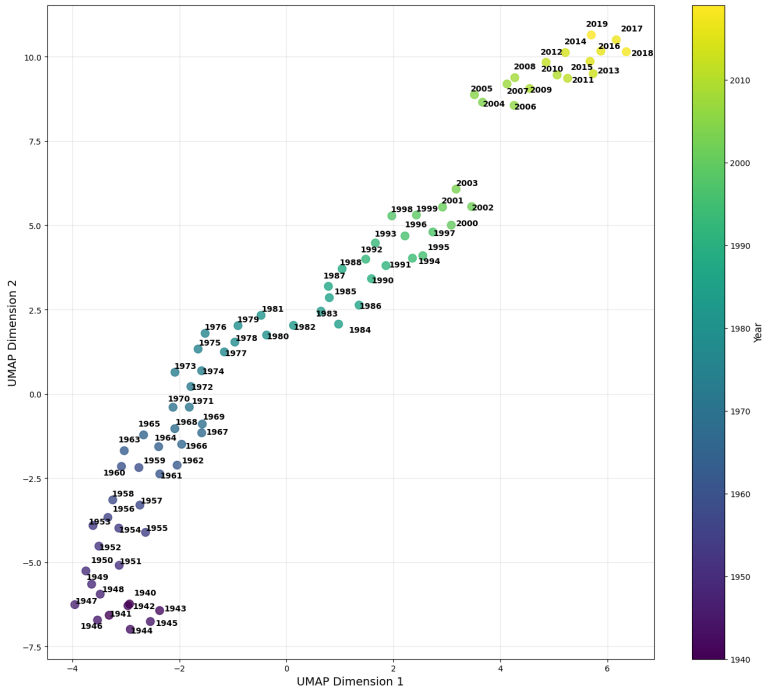

Figure 1: UMAP visualization of ASR paper embedding centroids from 1940-2020.

of published papers shift from decade to decade. Here I set n_neighbors as 25 and min_dist as 0.1 to yield tighter clusters for a larger number of embeddings. To further look into the internal dynamics of semantic change, I also explored and visualized the similarity between the semantics of a number of highly cited papers and different years' embedding centroids to see how the meanings of important works shift in relation to different years' general semantic meanings. All the methods above are applied to AJS and ASR separately to prevent the influence of potential differences in the writing styles between the two journals on my results.

4 Results

Figure 1 and Figure 2 respectively show the two-dimensional UMAP visualization of the centroids of paper embeddings in all given years from 1940 to 2020 for ASR and AJS. Each point denotes that year's embedding centroid with a color scale indicating the general chronological progression. Since UMAP preserves both global and local neighborhood structures, if two points are closer together than the other two points, their discursive or semantic meanings are supposed to be more similar to each other. From Figures 1 and 2, it then can be seen that any two successive years have greater semantic similarity than two randomly chosen years drawn from different decades. Moreover, embedding centroids in

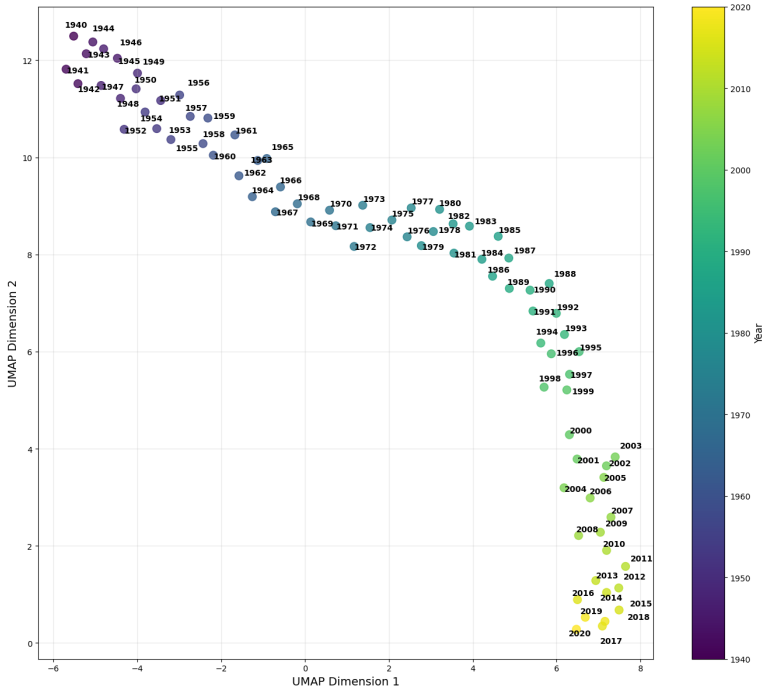

Figure 2: UMAP visualization of AJS paper embedding centroids from 1940-2020.

a relatively short timeframe between two to ten years tend to cluster together. Although such clustering has not been displayed in these two UMAP visualizations, it is a meaningful point to have a test on, which I will show with hierarchical clustering in the next section. It should also be noticed that within these small clusters of embedding centroids that span across a few years, they do not chronologically spread towards a clear direction on the UMAP visualizations. For example, the embedding centroids in the 1940s' cluster at the bottom-left corner in Figure 1 tends to be randomly spreading with no clear chronological pattern. However, it does not necessarily mean that they do not clearly develop towards a specific direction. In addition to the decade-level observation, at the macro level, it can be seen that all years' embedding centroids form a clear continuous curve with no isolated clusters nor abrupt leaps except the last two decades for ASR in Figure 1. This might indicate that the semantic meanings of different years' embedding centroids progressively differ from those in the 1940s. To test these ideas, I further computed the Euclidean and cosine distances between different years' embedding centroids to see if there is a clear pattern of drift.

To better examine if different years' embedding centroids in a close timeframe tend to cluster together, I applied hierarchical clusters to them. Figures 3 and 4 show that for both journals, yearly embedding centroids within the same period of time tend to be grouped in the same cluster. In Figure 3, at a relatively high level of the hierarchy, the embedding

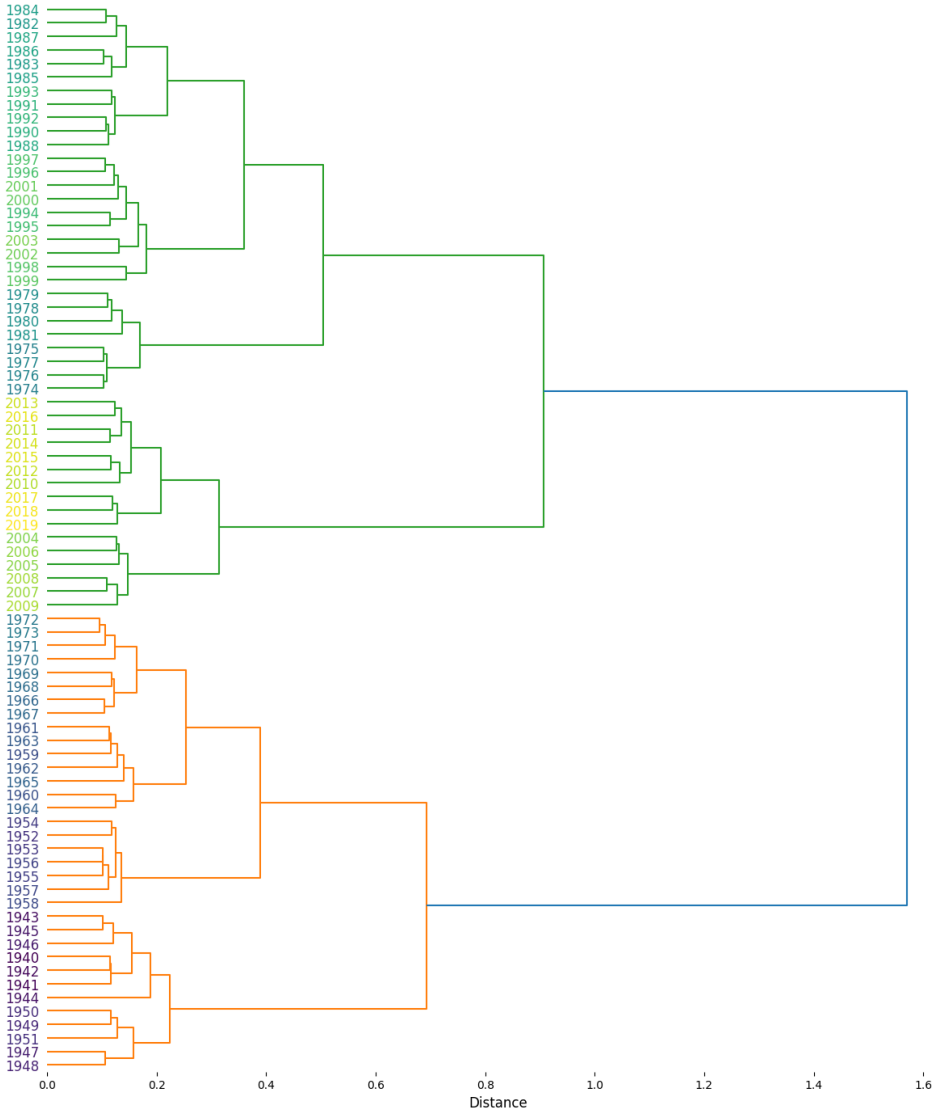

Figure 3: Hierarchical clustering of ASR paper embedding centroids.

centroids from ASR tend to form 4 clusters, each corresponding to 4 successive periods of time: 1940 to 1951, 1952 to 1973, 1974 to 2003, and 2004 to 2020. AJS in Figure 4 shares the same number of clusters: 1940 to 1961, 1962 to 1974, 1975 to 1999, and 2000 to 2020. Such a clear chronological clustering pattern holds true at lower hierarchies for both AJS and ASR. If you look closer at the hierarchical clustering, embedding centroids spanning three to five years tend to be grouped in low-level clusters. However, this does not hold true at the lowest level of the hierarchy. For example, rather than grouping 1983 and 1984 together, the upper bound of hierarchical clustering in Figure 3 clusters 1982 and 1984 together, indicating that these two years might have more semantics affinity than 1983 and

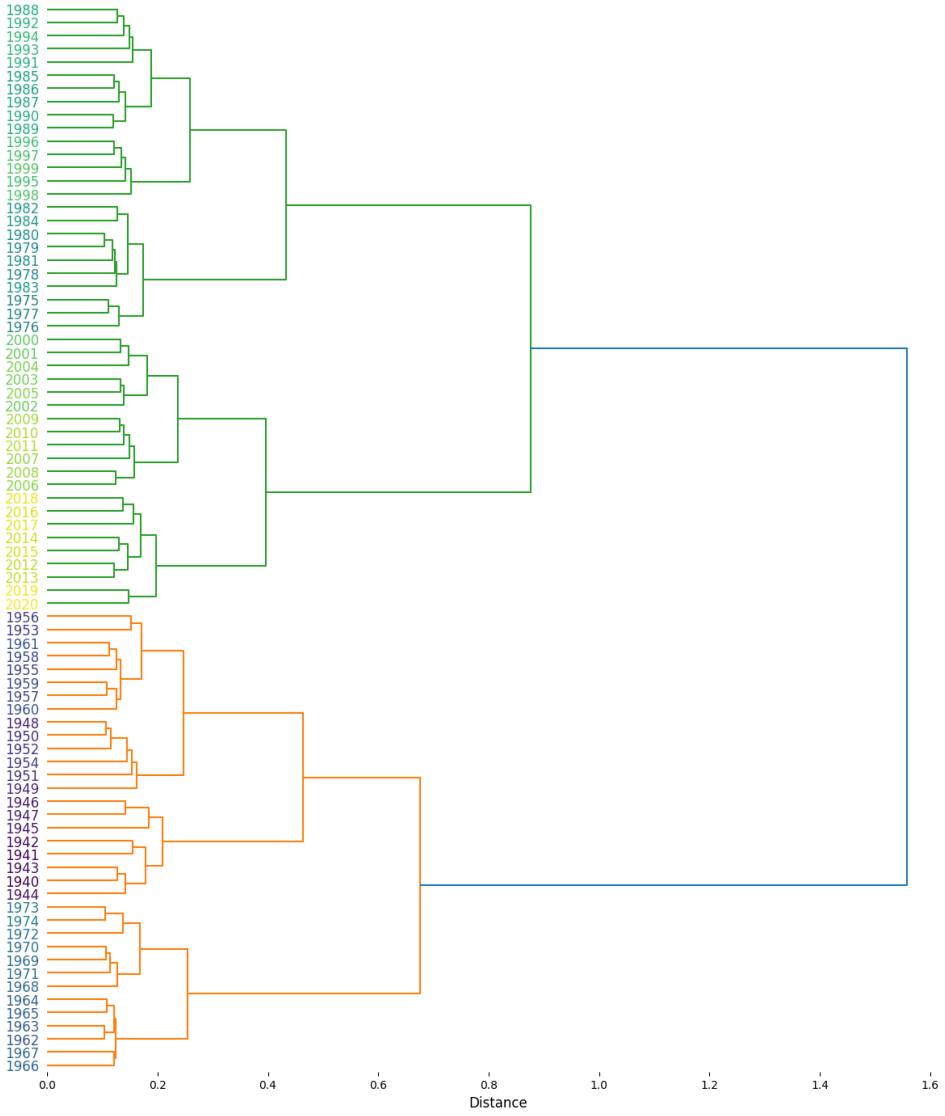

Figure 4: Hierarchical clustering of AJS paper embedding centroids.

1984. Nevertheless, the general chronological clustering pattern suggests that studies within the same era tend to share greater similarities with one another than those conducted in other periods. This observation can be further confirmed by the heat maps of year-to-year cosine similarities between the embedding centroids. As the two heat maps from Figures 5 and 6 demonstrate, yearly embedding centroids closer to each other in time tend to be redder than those from two different eras whose low similarities are depicted in blue. This confirms the conclusion drawn from the hierarchical clustering that papers' meanings from the same period of time have greater similarities with one another than those that are distant in time.

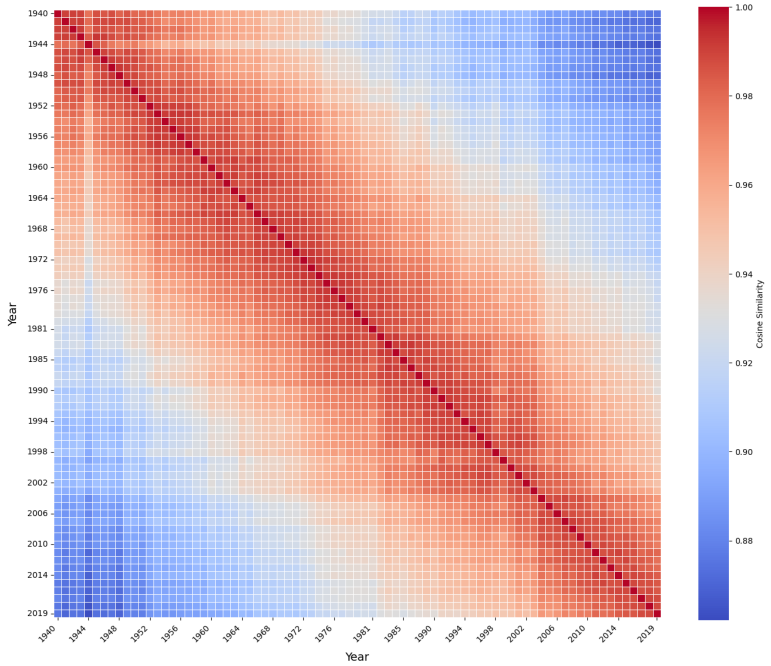

Figure 5: Heat map of year-to-year cosine similarity between ASR embedding centroids.

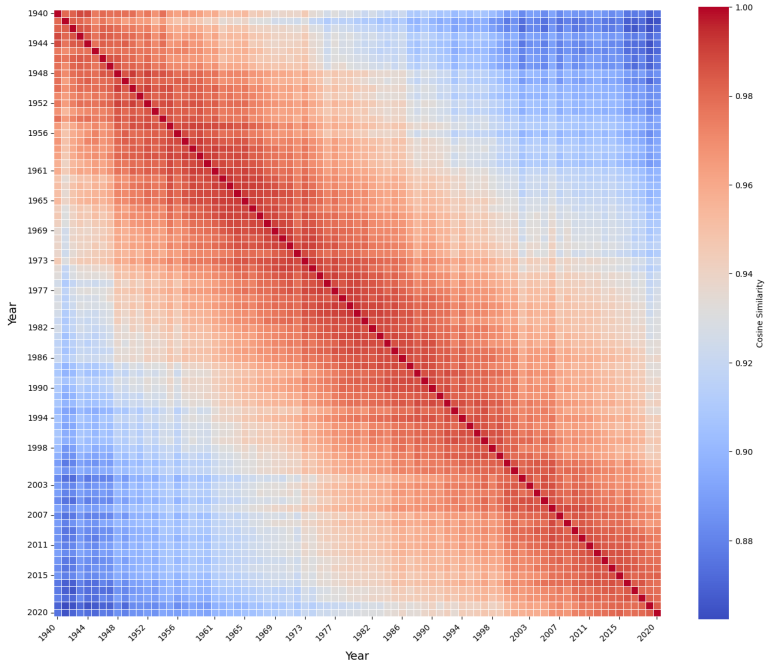

Figure 6: Heat map of year-to-year cosine similarity between AJS embedding centroids.

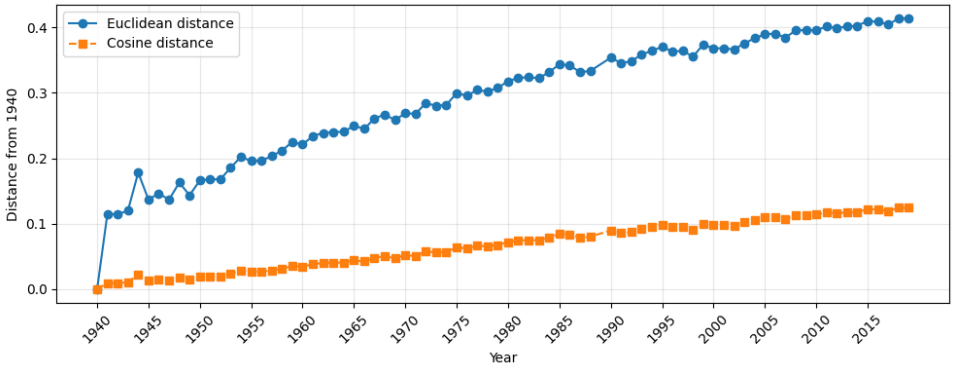

Figure 7: Distance from origin (1940) for ASR embedding centroids.

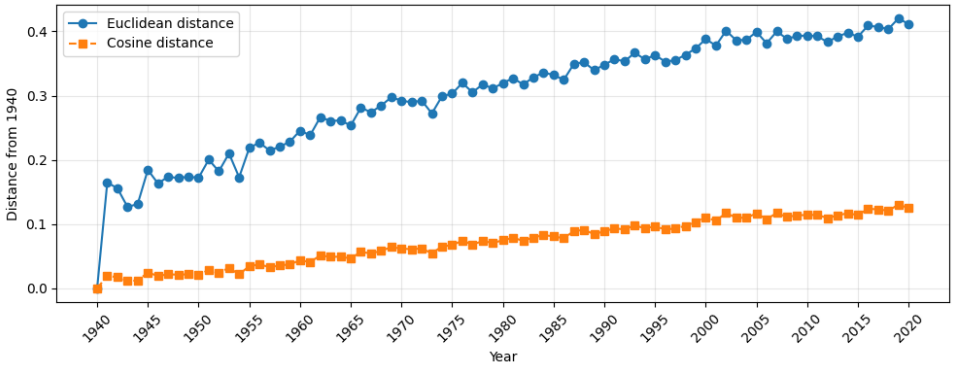

Figure 8: Distance from origin (1940) for AJS embedding centroids.

To further examine the semantic development of both journals since the 1940s, I computed the Euclidean and cosine similarity between the embedding centroid of the year 1940 and all subsequent years between 1941 and 2020 to show if there is a clear pattern of drift. I name this curve as distance-from-origin curve. Figures 7 and 8 show that the semantics of later years' research tend to incrementally and progressively deviate from the 1940s' research without sudden disruptions. Two things need to be emphasized here. First, the meanings in later years only gradually drifted away from the 1940s without a clear return. Second, each year's change from the previous year tends to be steady and stable without clear spikes or sharp downturns. To further look into the degree of semantic shift each year, I computed the Euclidean and cosine distances between each two successive years' embedding centroids and plotted them with Euclidean distance as the x-axis and cosine distance as the y-axis. Figures 9 and 10 show that there's a clear positive correlation between the two distances. Some years have experienced a greater degree of change in relation to previous years. For example, for ASR in Figure 9, the year 1944, and the years around 2000 experienced more dramatic changes than the others. For AJS in Figure 10, the years 1945, 1946, and those

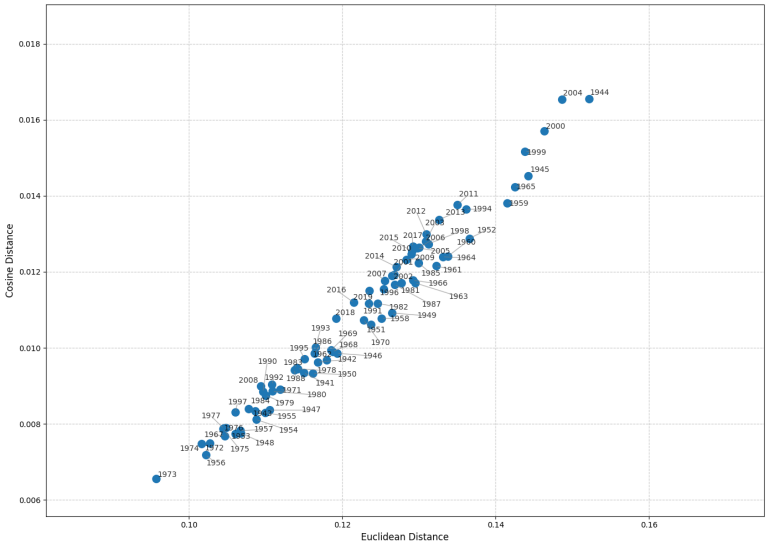

Figure 9: Correlation between Euclidean and cosine distances for successive years in ASR.

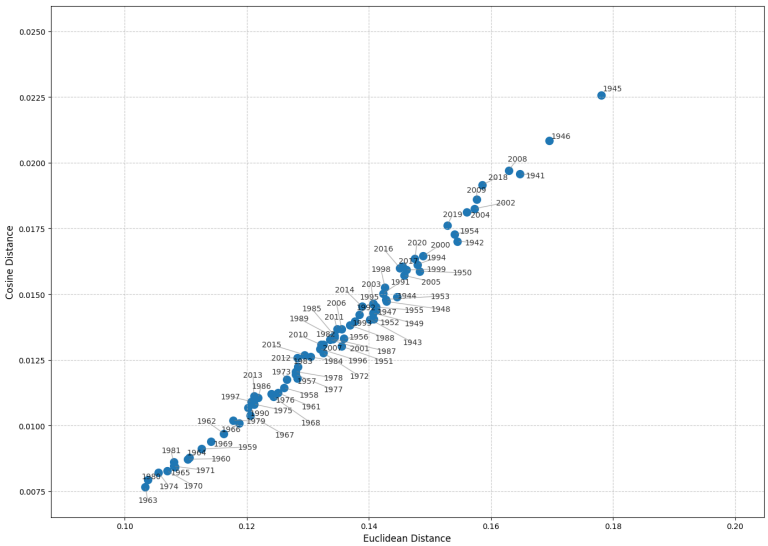

Figure 10: Correlation between Euclidean and cosine distances for successive years in AJS.

years in the 2000s stand out as years of sharp changes. However, there are no clearly separated groups of years distinct from the rest that indicate revolutionary or stagnant periods of meaning change. Such a pattern of change suggests that the underlying semantic drifts for both journals approximate a linear, incremental, and no-return fashion over time.

In addition to examining the embedding centroids of different years, it is also of vital importance to examine all the paper embeddings for ASR and AJS. Figures 11 and 12 are two UMAP visualizations of all the paper embeddings from ASR and AJS from the 1940s

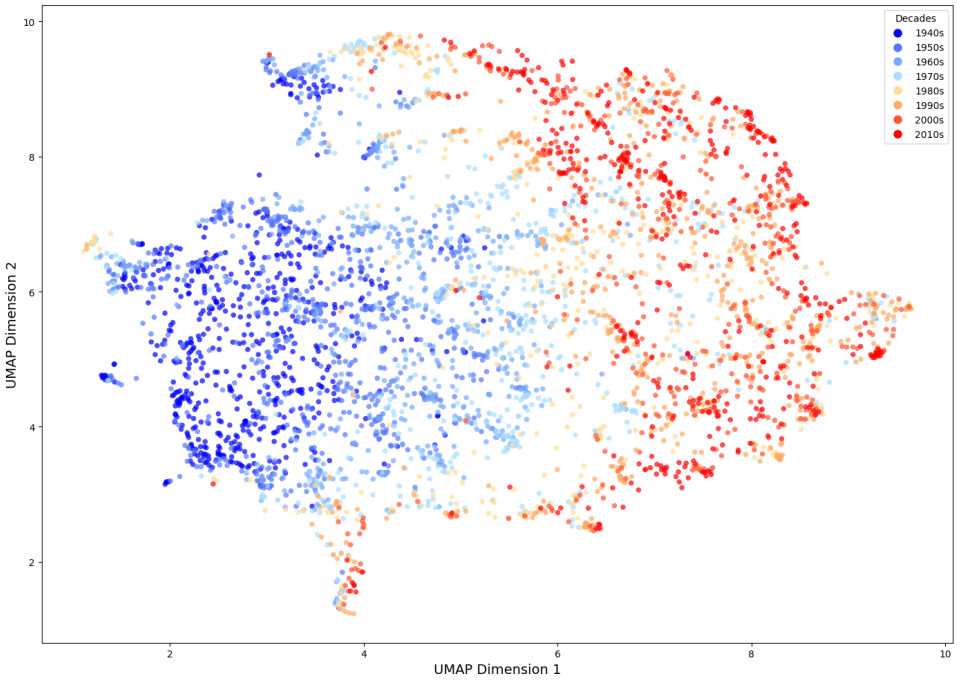

Figure 11: UMAP visualization of all ASR paper embeddings by decade (1940s-2010s).

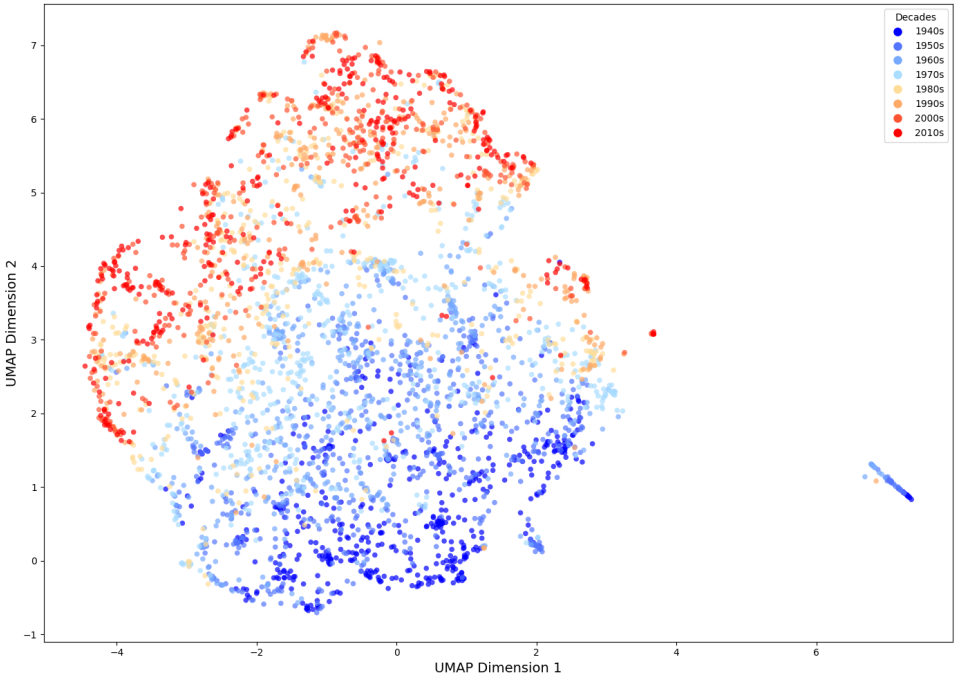

Figure 12: UMAP visualization of all AJS paper embeddings by decade (1940s-2010s).

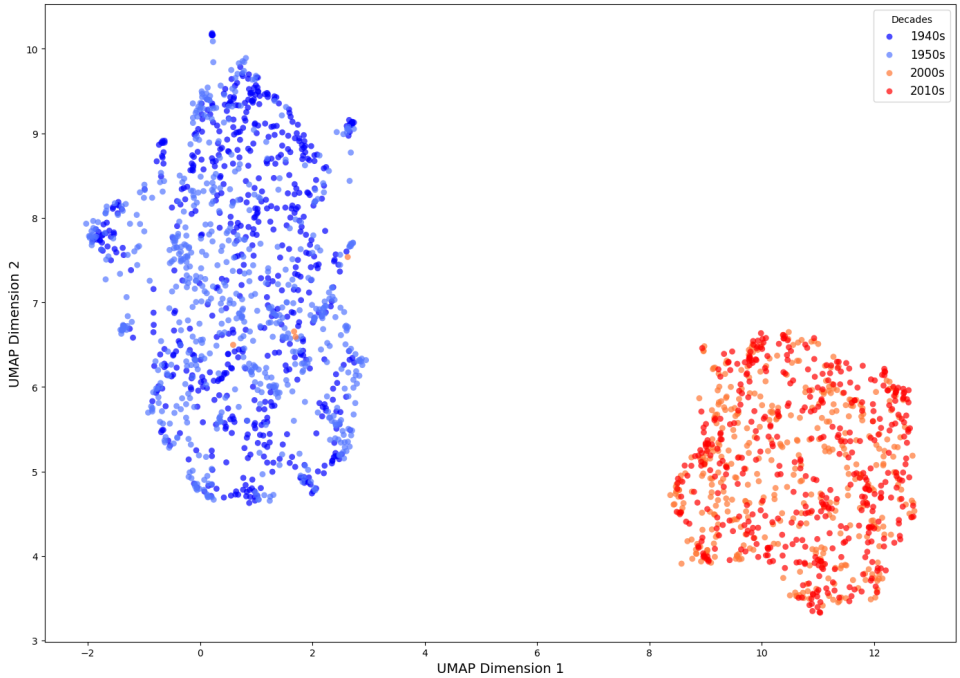

Figure 13: UMAP comparison of ASR paper embeddings between early decades (1940s-1950s) and recent decades (2000s-2010s).

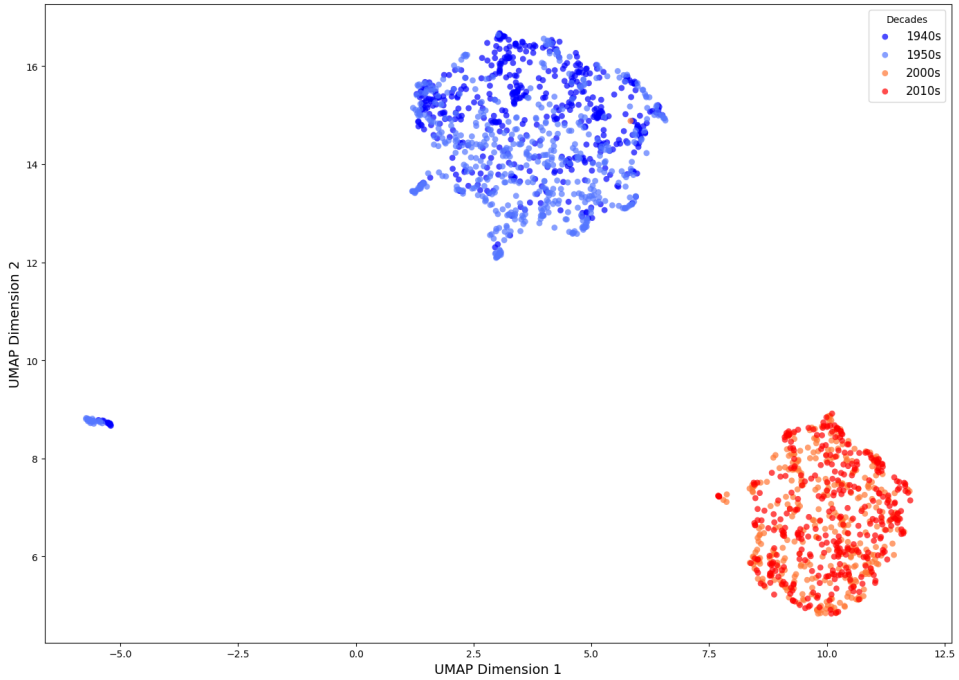

Figure 14: UMAP comparison of AJS paper embeddings between early decades (1940s-1950s) and recent decades (2000s-2010s).

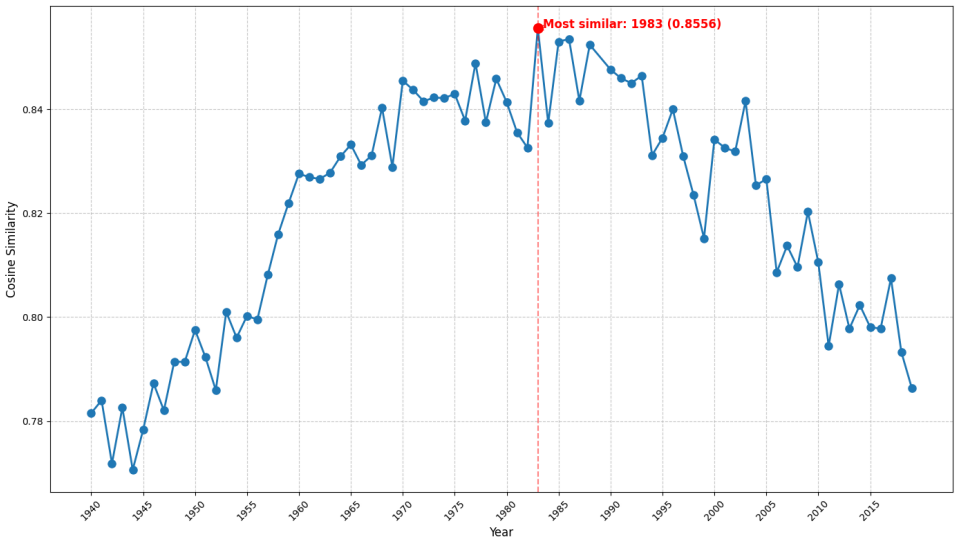

Figure 15: Semantic similarity between DiMaggio & Powell (1983) and ASR yearly embedding centroids.

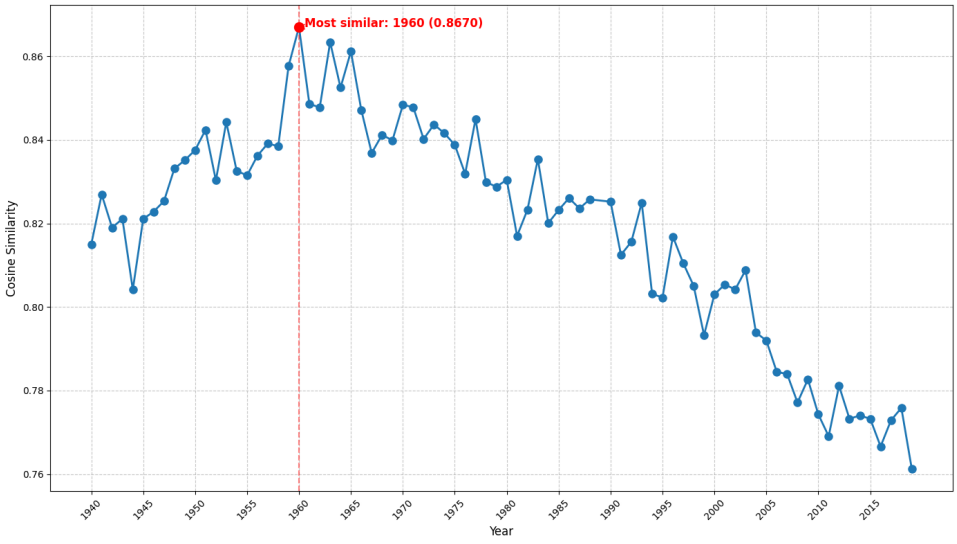

Figure 16: Semantic similarity between Gouldner (1960) and ASR yearly embedding centroids.

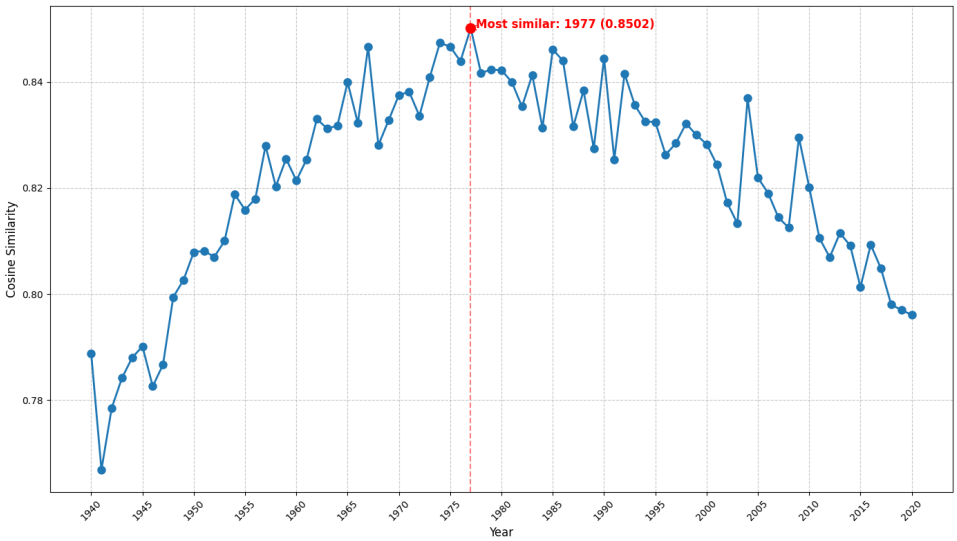

Figure 17: Semantic similarity between Granovetter (1973) and AJS yearly embedding centroids.

to 2010s. I split them by the eight decades and colored the embeddings with their decade membership. Several observations can be drawn from the two figures. First, the semantics of each decade's research tend to be clustered within a defined meaning space. While in previous decades such as the 1940s and 1950s papers' meanings tend to be clustered more closely with each other, those in the 2000s and 2010s tend to be more spread out. However, most of the embedding points for each decade are still confined in well-defined semantic spaces. Second, embeddings of two to three successive decades tend to have much overlap with each other. However, as time advances, they tend to overlap less and less with previous decades in the meaning space. This gradual meaning separation has a clear direction on the two UMAP visualizations, for Figure 11 from right to left and for Figure 12 from bottom to top. Third, for embeddings of decades that are far from each other in time, it can be seen that they have almost no shared meaning with each other. The embedding points in the 2000s and 2010s are almost all clustered within their own spaces, with only a very small number of them spreading into the semantic spaces of those in the 1940s and 1950s. This separation is better illustrated by Figures 13 and 14, in which the clusters of the publications in the 1940s and 1950s are dramatically separated from those in the 2000s and 2010s in the meaning space for both ASR and AJS. This pattern of semantic separation suggests that there is a significant semantic rupture between the early and current writings of sociology.

Furthermore, I singled out several key sociological papers that have high amount of citations and computed the cosine similarities between their embeddings and the yearly embedding centroids. Specifically, for ASR, I chose DiMaggio and Powell's *The Iron Cage*

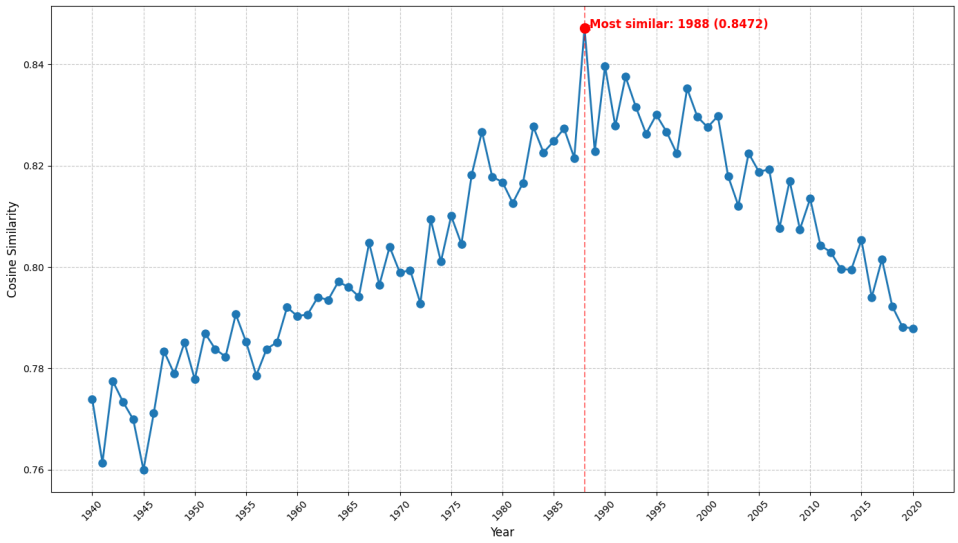

Figure 18: Semantic similarity between Coleman (1988) and AJS yearly embedding centroids.

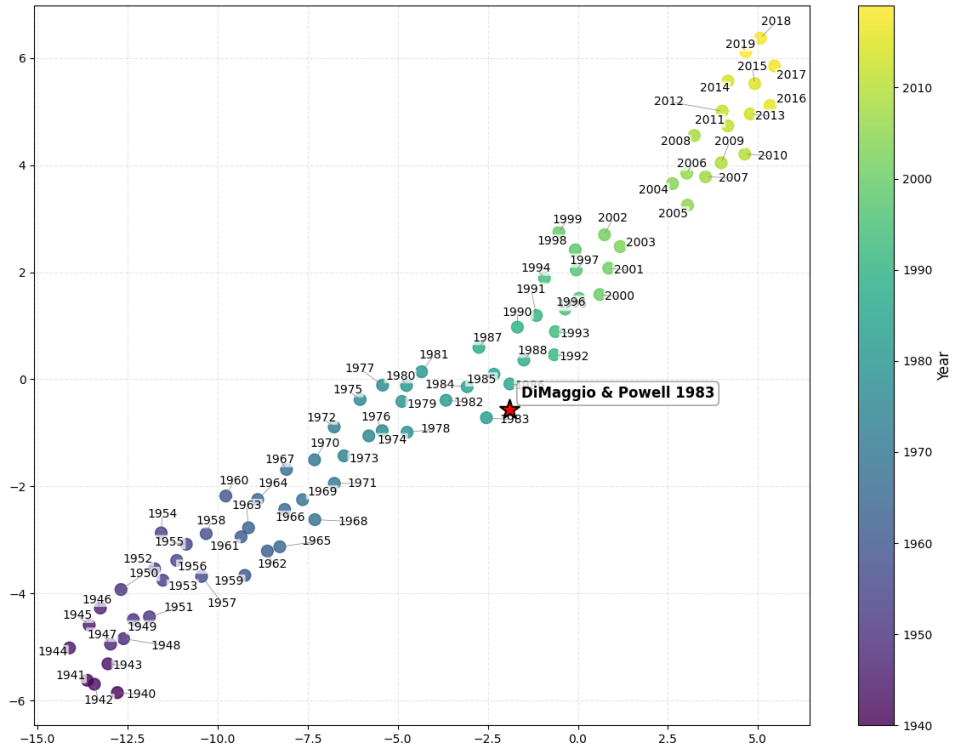

Figure 19: UMAP visualization of DiMaggio & Powell (1983) and ASR yearly embedding centroids.

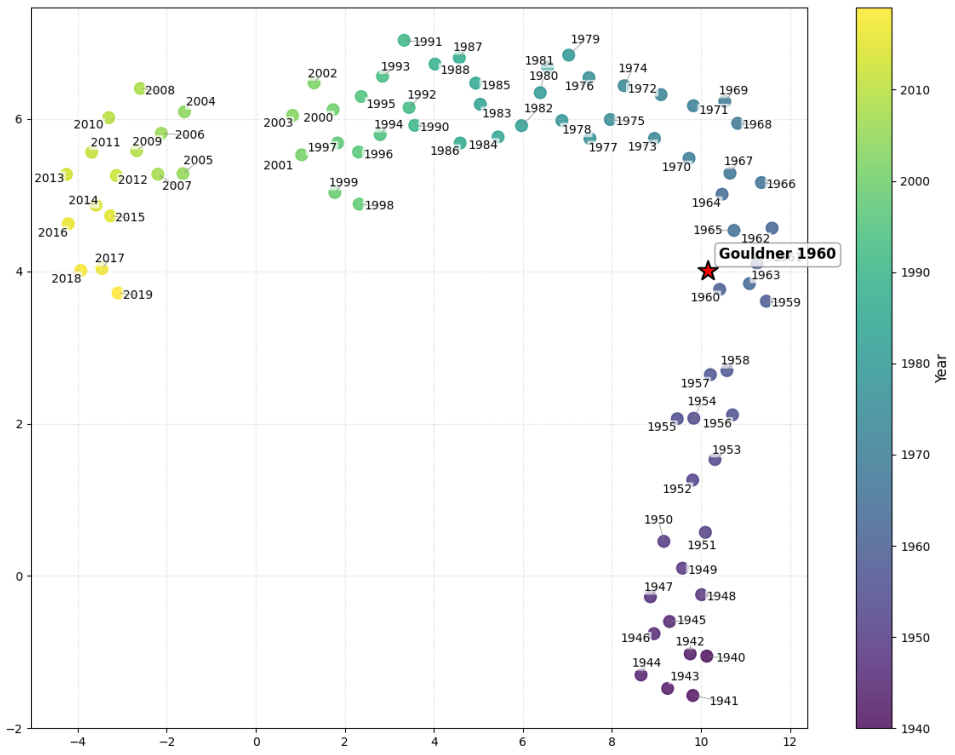

Figure 20: UMAP visualization of Gouldner (1960) and ASR yearly embedding centroids.

Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields in 1983 and Gouldner's *The Norm of Reciprocity: A Preliminary Statement* in 1960. For AJS, I chose Granovetter's *The Strength of Weak Ties* in 1973 and Coleman's *Social Capital in the Creation of Human Capital* in 1988. Their respective graphs are shown in Figures 15 to 18. Moreover, I also applied UMAP on every single paper and the yearly embedding centroids to visualize the relative semantic relationships between them. These visualizations are plotted in Figures 19 to 22. In general, the chosen papers shared a similar pattern of relation with the yearly embedding centroids. First, the semantic content of the selected papers shows only weak alignment with the overall semantics of years that are many years before or after their publishing date. Across Figures 15 to 18, early years' semantics gradually align more closely with the chosen papers' semantics, and after the period in which the papers were published, their semantic similarities gradually decline. Second, the year embedding centroids can successfully predict the periods if not the exact years in which those important works are published. In other words, the peak in semantic similarity between a selected paper's embedding and the yearly embedding centroids typically occurs in the very year the paper was published. Such a relationship could also be observed in the UMAP visualizations in Figures 19 to 22. The point representing the paper often lies in close proximity to the

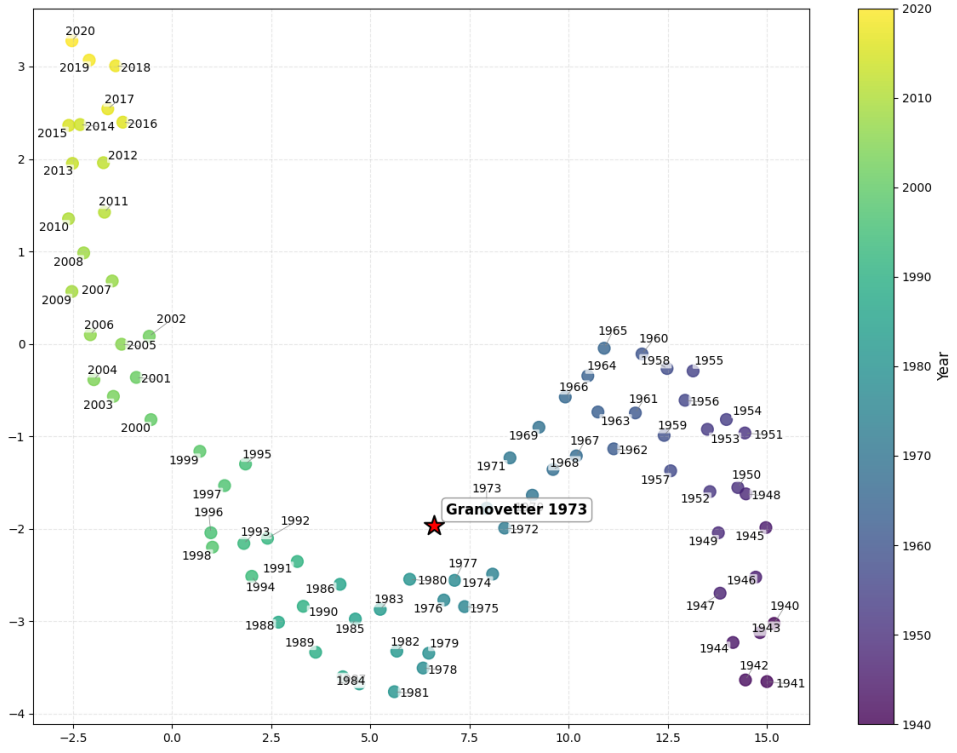

Figure 21: UMAP visualization of Granovetter (1973) and AJS yearly embedding centroids.

point corresponding to its year of publication, indicating a strong semantic alignment. For example, in Figure 15, DiMaggio and Powell's article shows the highest semantic similarity with the embedding centroid of 1983, the year it was published. Similarly, in Figure 19, the point representing their article on the UMAP is in closest proximity to the semantic meaning of 1983. Third, around the paper's publication year, there tends to be a span of several years during which the paper-to-year similarity remains relatively high. It can be named as a similarity plateau. Sometimes it only lasts for less than ten years as in the case of Gouldner's paper. However, in other cases, such plateaus would continue for two to three decades. For example, in the case of Granovetter's landmark paper on weak ties, this similarity plateau lasted from approximately 1965 to 1990 if we take an average similarity score of 0.84 as the plateau line. This dynamic semantic similarity between the chosen major papers and the yearly embedding centroids suggests that influential sociology papers published by the two journals maintain high semantic relevance with their surrounding literature for approximately one to two decades, after which their semantic alignment would gradually diminish.

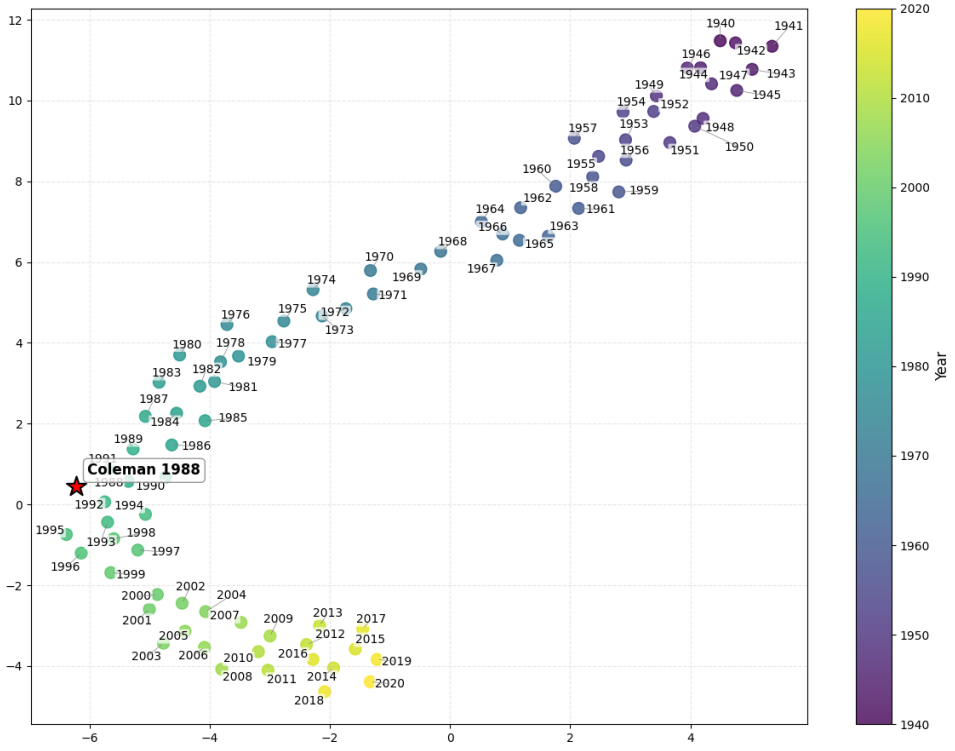

Figure 22: UMAP visualization of Coleman (1988) and AJS yearly embedding centroids.

5 Discussion

From the results above, it can be first concluded that if we take some complete key sociological corpora as the target of examination, the grand periods of sociology as claimed by some scholars disappeared. To be sure, sociological papers published closer in time tend to have greater semantic similarity with each other than those published farther. As the hierarchical clustering and heat maps in Figures 3 to 6 show, in a timeframe of three to five years, the general semantic meanings of different years' published papers have no clear differences. Yet beyond that timeframe, the semantic meanings of different years are more likely to be more similar or to be grouped in the same cluster if they are closer to each other in terms of time. This suggests that research published in closer years tends to share similar words, topics, and discursive forms. With more than 3,000 dimensions, these embeddings are supposed to capture even more abstract information for the papers, such as theoretical perspective or research framework. However, I would restrain myself from having such conclusions since it requires a deeper and more thorough examination of the texts.

However, even though adjacent years show semantic similarity, I find no clear evidence for distinct research periods. A basic assumption here is that in order for a distinct research period to emerge, there should be years of semantic rupture that signal a transitional or

revolutionary stage that divides two research periods. However, as the Euclidean and cosine distances in Figures 7 and 8 show, the differences in general semantic meanings between the years after 1940 and the year 1940 widen over time in an almost linear fashion without a dramatic increase or decline in the meaning difference. Such a smooth and steady pattern of semantic change can also be observed in Figures 11 and 12. Despite the positive correlation between the Euclidean and cosine distances, there is no clear cluster of both high Euclidean and cosine distances at the right upper corner of the plots that may potentially signify a clear revolutionary change in the semantic meanings of the published papers. Such a gradual and monotonic semantic shift suggests that instead of changing through sharply delineated periods, sociological scholarship evolves in a gradual and steady manner. As such, the first hypothesis is confirmed.

Second, from Figures 15 to 18, it can be seen that sociology is characterized by multiple several-year subprocesses that are initiated or involved by impactful works. The semantic meanings of the four chosen papers in the four figures have a considerable timeframe in which their relevance with those years' general semantic meanings is relatively high. This indicates that these highly cited papers may invite more papers that are semantically similar to them in the next years or be involved in discussions related to their topics that have lasted for a while years. Such a relevant timeframe for the chosen impactful works lasts from a few years to two to three decades. After that, the relevance would gradually decline without a clear return. Such a semantic trend suggests two things. First, an important piece of work in sociology would invite a lasting discussion or emerge in an academic discussion as it progressed. Second, such an academic conversation would gradually lose its popularity in the later years of research. This suggests that the relevance of impactful sociological work for the papers published after them cannot be everlasting. The focus of the discussion would migrate from the topics or the issues they are concerned with to other fields. It is these various micro-level academic discussions that are time-limited that constitute the general landscape of sociology. As such, the second hypothesis is confirmed.

Third, as Figures 11 to 14 suggest, there is a significant rupture in the semantic meanings between the early years of sociological writings and the current stage. Figures 11 and 12 reveal that the change in semantic meanings of the published papers throughout the years is to an extent directional. The semantic spaces of later decades gradually detached from the spaces of the earlier decades through a monotonic path until they almost have no overlap with each other. Such a trend is clearly mapped in Figures 13 and 14 in which the semantic spaces representing the 1940s to 1950s and 2000s to 2010s are distinctly separated from each other with only a few embedding points in the 2000s and 2010s joining the 1940s and 1950s cluster. This clear separation suggests that sociology in the new century only has a very weak semantic relationship with those in the 1940s and 1950s. To an extent, it has

forgotten its past, descending into the oblivion of its remote history. As such, the third hypothesis is confirmed.

6 Conclusion

In conclusion, by applying text embeddings on papers published in the American Sociological Review and the American Journal of Sociology from 1940 to 2020, this paper finds that while the meanings of sociological papers tend to gradually and incrementally shift away from those in the 1940s, those papers published in recent two decades tend to have a semantic rupture from those published in early years. Such a gradual shift of meaning consists of the rise and fall of multiple micro-level research discussions. Such conclusions support previous claims that sociology has been increasingly losing sight of its past (e.g., Turner, 1989; Stinchcombe, 1994). Furthermore, they add the nuance that this oblivion is not the consequence of an abrupt event but a gradual and lengthy historical process.

The current research has several issues. First, the text embeddings it used were not highly interpretable. Compared with word embeddings, where the semantic relationships between different concepts can be directly examined, document-level embeddings tend to function as opaque high-dimensional representations. It is therefore difficult to interpret the results and determine which specific elements are accountable for the change in general meanings. It also complicates the validation of the results. For this reason, I tend to refrain from offering detailed interpretations of what exactly the semantic similarity captures in addition to words and topics. Second, the examination of the second hypothesis is relatively inadequate. More examples of papers should be used to examine the relationships between individual papers' meanings and the general dynamics of papers' meanings in time. Third, this paper only examines two top American sociological journals, potentially reducing the generalizability of the results for sociology. Other potential resources can be used here such as books and sociological journals in Europe and the global south to expand the generalizability of the research.

Nevertheless, this paper provides a helpful heuristic for comparing the semantics of texts over time. By averaging all papers' embeddings in a given year, it enables the comparison of the meanings of texts from year to year. Similar stories might be unfolding in other areas of the social sciences or even beyond them. Further research could apply the methods proposed here to develop digital genealogies in various disciplines, tracing how the meanings of texts in academic or literary fields evolve over time.

Data and Code Availability Statement

The data and code used in this study are publicly available at this link. All analyses were conducted using Python, and the scripts necessary to reproduce the figures and results are included in the repository. Additional information is available upon request.

References

Abbott, A. D. (2001). *Chaos of disciplines*. University of Chicago Press.

Aceves, P., & Evans, J. A. (2024). Mobilizing Conceptual Spaces: How Word Embedding Models Can Inform Measurement and Theory Within Organization Science. *Organization Science*, *35*(3), 788–814. https://doi.org/10.1287/orsc.2023.1686

Alexander, J. C. (2014). *Theoretical logic in sociology*. Routledge.

Arseniev-Koehler, A., Cochran, S. D., Mays, V. M., Chang, K.-W., & Foster, J. G. (2022). Integrating topic modeling and word embedding to characterize violent deaths [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *119*(10). https://doi.org/10.1073/pnas.2108801119

Bestvater, S. E., & Monroe, B. L. (2023). Sentiment is Not Stance: Target-Aware Opinion Classification for Political Text Analysis [Publisher: Cambridge University Press (CUP)]. *Political Analysis*, *31*(2), 235–256. https://doi.org/10.1017/pan.2022.10

Bourdieu, P., & Bourdieu, P. (2002). *Distinction: A social critique of the judgement of taste* (11. print). Harvard Univ. Press.

Coleman, J. S. (1988). Social Capital in the Creation of Human Capital [Publisher: University of Chicago Press]. *American Journal of Sociology*, *94*, S95–S120. https://doi.org/10.1086/228943

DiMaggio, P. J., & Powell, W. W. (1983). The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields [Publisher: SAGE Publications]. *American Sociological Review*, *48*(2), 147. https://doi.org/10.2307/2095101

Giddens, A. (1986). *The Constitution of Society: Outline of the theory of structuration* (First paperback edition). University of California Press.
Bibliographie: Seite [379]-391.

Gouldner, A. W. (1960). The Norm of Reciprocity: A Preliminary Statement [Publisher: SAGE Publications]. *American Sociological Review*, *25*(2), 161. https://doi.org/10.2307/2092623

Granovetter, M. S. (1973). The Strength of Weak Ties [Publisher: University of Chicago Press]. *American Journal of Sociology*, *78*(6), 1360–1380. https://doi.org/10.1086/225469

Habermas, J., McCarthy, T. A., & Habermas, J. (2007). *Reason and the rationalization of society* (Nachdr.). Beacon.

Joas, H., & Knöbl, W. (2009, July). *Social Theory: Twenty Introductory Lectures* (A. Skinner, Trans.; 1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139878432

Keith, B., & Ender, M. G. (2004). The Sociological Core: Conceptual Patterns and Idiosyncrasies in the Structure and Content of Introductory Sociology Textbooks, 1940-2000. *Teaching Sociology*, *32*(1), 19–36. https://doi.org/10.1177/0092055X0403200102

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review*, *84*(5), 905–949. https://doi.org/10.1177/0003122419877135

Licht, H. (2023). Cross-Lingual Classification of Political Texts Using Multilingual Sentence Embeddings [Publisher: Cambridge University Press (CUP)]. *Political Analysis*, *31*(3), 366–379. https://doi.org/10.1017/pan.2022.29

Lin, G. (2025). Using cross-encoders to measure the similarity of short texts in political science [Publisher: Wiley]. *American Journal of Political Science*. https://doi.org/10.1111/ajps.12956

Meyer, J. W., & Rowan, B. (1977). Institutionalized Organizations: Formal Structure as Myth and Ceremony [Publisher: University of Chicago Press]. *American Journal of Sociology*, *83*(2), 340–363. https://doi.org/10.1086/226550

Nelson, L. K. (2021). Leveraging the alignment between machine learning and intersectionality: Using word embeddings to measure intersectional experiences of the nineteenth century U.S. South. *Poetics*, *88*, 101539. https://doi.org/10.1016/j.poetic.2021.101539

Peng, H., Ke, Q., Budak, C., Romero, D. M., & Ahn, Y.-Y. (2021). Neural embeddings of scholarly periodicals reveal complex disciplinary organizations. *Science Advances*, *7*(17), eabb9004. https://doi.org/10.1126/sciadv.abb9004

Starr, J. M. (1983). Specialization and the development of sociology: Differentiation of fragmentation? *Qualitative Sociology*, *6*(1), 66–86. https://doi.org/10.1007/BF00987198

Stinchcombe, A. L. (1994). Disintegrated Disciplines and the Future of Sociology. *Sociological Forum*, *9*(2), 279–291. Retrieved April 2, 2025, from https://www.jstor.org/stable/685046?seq=1

Susen, S., & Turner, B. S. (2011). Tradition and innovation in classical sociology: Tenth Anniversary Report of*JCS* [Publisher: SAGE Publications]. *Journal of Classical Sociology*, *11*(1), 5–13. https://doi.org/10.1177/1468795x10391451

Szelenyi, I. (2015, April). The triple crisis of sociology. Retrieved November 1, 2024, from https://contexts.org/blog/the-triple-crisis-of-sociology/

Taylor, M. A., & Stoltz, D. S. (2021). Integrating semantic directions with concept mover's distance to measure binary concept engagement. *Journal of Computational Social Science*, *4*(1), 231–242. https://doi.org/10.1007/s42001-020-00075-8

Turner, J. (1989). The Disintegration of American Sociology: Pacific Sociological Association 1988 Presidential Address [Publisher: SAGE Publications]. *Sociological Perspectives*, *32*(4), 419–433. https://doi.org/10.2307/1389130

Turner, J. (2001). Introduction – The Fragmentation of Sociology [Publisher: SAGE Publications]. *Journal of Classical Sociology*, *1*(1), 5–12. https://doi.org/10.1177/14687950122232431

Vicinanza, P., Goldberg, A., & Srivastava, S. B. (2023). A deep-learning model of prescient ideas demonstrates that they emerge from the periphery (J. Van Bavel, Ed.) [Publisher: Oxford University Press (OUP)]. *PNAS Nexus*, *2*(1). https://doi.org/10.1093/pnasnexus/pgac275