# Meaning, Understanding, and Quantification in the History of Ideas

Jeff Jacobs
jjacobs3@cs.stanford.edu

April 22, 2023

**Abstract**

What are political thinkers *doing* with their words when they write a text, engage in a debate, or give a speech? We propose a "computational political theory", pairing recent breakthroughs in computational linguistics with the hermeneutic practices of intellectual history, as a set of tools for mapping out the political-discursive fields within which ideas circulate. We show, via a series of historical case studies, how a particular class of computational-linguistic algorithms called word embeddings are able to capture subtle differences in how authors employ certain contested terms (liberty, freedom, sovereignty, etc.) by explicitly modeling both the words and the contexts they're used in across a corpus of texts. We argue that these context-sensitive language models thus represent powerful and underutilized tools for historical research, and provide a framework for their use in developing, testing, and revising our understandings of key questions in the history of political thought.

## Contents

# 1 Introduction

What are political thinkers *doing* with their words when they perform a political speech act—when they publish a treatise, give a speech, argue for a piece of legislation, and so on? Since the 1960s a group of historians centered around Cambridge University professors Quentin Skinner, J. G. A. Pocock, and John Dunn have changed the way we understand several key political thinkers and texts, by consciously developing and applying a linguistic-philosophical and *context-sensitive* methodology which places this question at the center of historical inquiry (Skinner 1969; Pocock 1985). As part of this endeavor the Cambridge School, as this group came to be known, placed a strong emphasis on the need for historians to explicitly describe and justify the methodologies

being employed in their studies[1] Most Cambridge School works therefore begin with extensive methodological introductions, in which particular modes of inquiry are introduced and justified prior to their application in the remainder of the work. Kenneth Minogue describes Skinner's early work, for example, as

> "primarily of interest to philosophers not for its excellent account of European thought about the state but for the self-conscious philosophy which has gone into it. It is a rare historian who pauses to get his philosophy in order before he embarks on a major enterprise" (Tully 1988, p. 176)

Thus the first of the three volumes of Quentin Skinner's *Visions of Politics*, for example, is entirely devoted to method (providing 209 pages of methodological justification before discussing the Renaissance and Thomas Hobbes in Volumes 2 and 3, respectively), while J. G. A. Pocock's *Virtue, Commerce, and History* begins with an intensive 36-page treatise on methodology, applying lessons learned from prior investigations to refine his approach before embarking on the inquiries of the remainder of the book.

It was not their emphasis on method *as such*, however, that set the Cambridge School apart from other historiographic schools—indeed, many other postwar hermeneutic approaches (e.g., Derridean deconstructivism) have also accumulated a vast methodological literature[2]. Rather, it was their focus on *contextual analysis* of texts—their call to de-emphasize the "so-called 'classic texts' [...] and focus instead on the more general social and intellectual matrix out of which [these] works arose" (Skinner 1978a, p. x)—that enabled their remarkable impact on political theory and the history of political thought, and that motivates our adoption and extension of their approach via computational-linguistic tools.

[1] Pocock, in fact, puts the point more strongly, arguing that the Cambridge School *introduced* this self-reflection into the field of intellectual history:

> "it was only in the middle 1960s, with the first appearance of writings by Quentin Skinner, that historians of political thought began to state the logic of their own inquiry and pursue it into fields where it encountered the philosophy of language." (Pocock 1985, p. 3)

We explore this claim—placing the Cambridge School itself within its broader historical context to argue that Karl Popper, more so than Skinner, can be credited with introducing this focus (a focus whose lineage contains important political consequences)—in Section 2.3 below.

[2] See Skinner 1990 or Tully 1988, for example, for conversations between Cambridge School practitioners and scholars representing a wide range of alternative hermeneutic perspectives.

In fact, as we will argue in the remainder of this chapter, these computational tools represent a radical breakthrough for historiographic practice, unlocking for the first time the full potential of the Cambridge School approach for enriching our understanding of the history of political thought. This immanent potential, we argue (building on the argument of London 2016), was latent in the pre-computational formulation of the Cambridge School approach, but unrealizable in practice due to the fundamental limits of human reading comprehension. Cambridge School studies to this point, notwithstanding the massive impact they've already had on the landscape of our historical knowledge, have been restricted to instances in which the relevant context necessary to understand a given thinker or text was on a scale small enough to allow one person (or a small team of people) to not only *read* and *process* these pieces of context, but also to then *synthesize* what had been learned into a coherent narrative or explanation. Hence Quentin Skinner explicitly describes his overall historical project as one restricted to studying "the European tradition only during its most formative days, which I take to have been the sixteenth and seventeenth centuries." (Scott and Keates 2001, p. 15) Cambridge School researchers have thus aptly demonstrated—in studying clerical disputes in the early stages of the Protestant Reformation (Skinner 1978b), for example, or the revival of Roman republican ideals in sixteenth century Florence (Pocock 1975)—the efficacy of this approach when the *scope* of the community of discourse under examination can be feasibly circumscribed.

These two examples, however, and other impactful Cambridge School works, are noteworthy perhaps for the very reason that they run up against the limits of human ability: they examine "general social and intellectual matri[ces]" capacious enough to derive profound insights into the history of political thought, but not so expansive as to render the project infeasible for one or a few people. Pocock summarizes his goal in *Virtue, Commerce, and History*, for example, as one of writing "a history of actors uttering and responding in a shared yet diverse linguistic context." Importantly, Pocock asserts, the subject of this history is bounded not only by this shared context but also by the fact that its actors communicate via the "internal" circulation of their writings, in the form of published manuscripts:

"We need not therefore apologize for the unrepresentative elitism of studying only those

readers whose responses were verbalized, recorded, and presented." (Pocock 1985, p. 18)

We argue that in general, but especially in the case of Marx and his desire to enact social change on a mass scale (a desire succinctly and powerfully captured in his 11th Thesis on Feuerbach[3], the very nature of the political discourse we hope to study was transformed over the course of the nineteenth century, in both its material and ideological dimensions. Materially, as we discuss in more detail below, the accelerating development of communication and transportation technology combined with the spread of literacy gave rise to international political-discursive fields which were no longer restricted to the small circle of "literary elites" referred to by Pocock (e.g., the eighteenth-century French *salonnières*), but now encompassed burgeoning mass publics engaging via new developments like workers' newspapers and study groups. These infrastructural and institutional changes were inevitably accompanied by ideological developments, in terms of e.g. the audiences that writers had in mind when composing and issuing political speech acts: writers like Marx and Engels who were interested in *mobilizing* these mass publics had to tailor both the content and style of their writings—references, metaphors, idiomatic expressions, etc.—in such a way as to appeal to a worker in Portugal just as well as one in Poland.

Indeed, while late eighteenth-century developments like the hortatory printings of Marat and other French Revolutionary pamphleteers represent early moves in this direction, their scope was limited mainly to Paris or other prominent regions of France. As we will argue below, the move from these parochial appeals to Marx and Engels' address to the "workers of the world" 60 years later nicely captures the magnitude of the expansion of scope we are referring to here.

This historical trajectory thus leads us to the question of why—i.e., on the basis of what epistemological principle—we should believe that the sufficiently-relevant context for understanding a given thinker or text can be restricted to this individually-manageable size, as more and more entrants contribute to these "shared yet diverse linguistic context[s]"?[4]. If a given historical event

[3]"Die Philosophen haben die Welt nur verschieden *interpretiert*, es kömmt drauf an sie zu *verändern*." (*Marx-Engels Gesamtausgabe*, IV/3, p. 21; *Marx-Engels Werke*, Vol. 3, p. 7) ("The philosophers have only interpreted the world in various ways; the point is to change it.") (*Marx-Engels Collected Works*, Vol. 5, p. 5)

[4]It is important to note at this point that an alternative approach to intellectual history, the *Begriffsgeschichte* approach of Reinhart Koselleck and others (mainly in the German-speaking world), does not avoid this drawback.

is complex enough, that is, if the collection of salient factors is so vast as to make an individual close-reading-based study truly impossible, then to conduct such a non-computational contextual study is to fall prey to the so-called drunkard's search fallacy: searching for your lost keys under a streetlight not because you dropped them near the streetlight but because the area around the streetlight is the only spot illuminated enough to see[5].

Even if we do believe that the relevant context can be restricted to this extent, however, there *are* important historical-epistemological principles (such as the covering-law model, or more general principles of replicability and Bayesian inference) which imply that the conclusions we draw from some individually-studied context are meaningful precisely to the extent that they are robust to the collection of additional evidence—in this case, evidence derived from the computational study of texts beyond the human-readable frontier. In other words, even if one rejects our stronger claim that these computational methods are *necessary* for conducting Cambridge School-style studies of nineteenth- and twentieth-century political thought, we hope to still present a compelling argument for their use in "checking one's work"—for example, verifying that nothing in the output flatly contradicts one's findings—after carrying out a non-computational study.

Although we maintain that even the historical *explananda* of the famous Cambridge School studies—the Renaissance, the Protestant Reformation, or Pocock's Machiavellian Moment—are already too complex for any individual to fully assimilate their relevant contexts, for the purposes of this work we rely on a less strong claim, one that does not preclude this possibility for cases of pre-industrial phenomena. Namely, we argue that by the time of the French Revolution, and especially by the time of the advent and spread of Marxism across Europe, literacy and communication

Though this approach centers *concepts* (*begriff*) as the unit of study, rather than thinkers or texts, within our framework this essentially amounts (we argue) to debating whether it is more fruitful to focus on the white squares or black squares of a chessboard if we want to understand chess: though Cambridge School practitioners start with individual thinkers or texts, they inevitably must engage with the concepts used by these thinkers or texts. Similarly, though *Begriffsgeschichte* analyses start with concepts, they inevitably must understand them by way of the individual thinkers and texts which shaped and altered (or failed to shape/alter) these concepts. We do note, however, that there is a more well-formed literature on the fruitful use of computational-linguistic tools for *Begriffsgeschichte* than there is for the Cambridge School approach. See, e.g., Wevers and Koolen 2020.

[5]The name of this observational bias refers to the following joke: A policeman sees a drunkard searching for something under a streetlight and asks what the drunkard is looking for. He says he's lost his keys, so the officer offers to help and joins him in searching for the keys around the streetlight. After a few minutes, the policeman asks him if he's sure he lost them here. The drunk replies: "no, I lost them in the park across the street". When the policeman angrily demands to know, then, why they are searching here, the drunkard calmly replies, "because this is where the light is!".

technology had expanded the scope of the relevant "social and intellectual matrix" underlying key political texts to the point that a non-computational assimilation of this matrix (or, in other words, its comprehension in totality by one individual) is no longer possible.

Indeed, as we move from the French Revolution to the 19th-century advent and spread of Marxism, we enter an era where the conduct of political polemics undergoes a massive set of changes. Not only are political speech acts increasingly conducted on a global scale, but also increasingly able to penetrate below the level of a society's elites, reaching and addressing (or failing to address) the concerns of ordinary citizens who found themselves swept up in the violent upheavals engendered by the Industrial Revolution, with its "constant revolutionising of production, uninterrupted disturbance of all social conditions, [and] everlasting uncertainty and agitation" (MECW 6487). The effect of these changes can be observed, for example, in the Preamble to the *Communist Manifesto*—an archetypical example of an illocutionary speech act in the form of a text—where it is explained that

> *Communists of various nationalities have assembled in London and sketched the following manifesto, to be published in the English, French, German, Italian, Flemish and Danish languages.* (*Marx-Engels Collected Works*, Vol. 6, p. 481)

Thus we see that, in the eyes of the Communist League who issued the manifesto in 1848, the audience for their polemic was not restricted to the workers of London, nor was it intended for the elites of England, France, Germany, Italy, Belgium, and Denmark, but instead aimed to engage *all* workers across *all* of these nations[6].

The increasingly global character of these polemics highlights another dimension of the intractability of the Cambridge School approach with respect to important historical phenomena, namely, that the researcher is forced to read not just a vast *number* of texts, but a vast number of texts across a vast range of *languages*. The challenge of explaining the 18th- and 19th-century ideas which emerged from the discourse around freedom among creoles of the colonies of North

[6]Indeed, as discussed in Feuer 1963 and Gourevitch 2015 (pp. 185–189), German exiles would spread these ideas beyond the borders of Europe in the aftermath of 1848–1849, transporting them across the Atlantic and bringing them to bear on discussions around the "labor republicanism" propounded by radical U.S. organizations like the Knights of Labor and (later) the Industrial Workers of the World (IWW).

and South America, for example—a challenge taken up by Joshua Simon in *The Ideology of Creole Revolution*—already brushes up against the limits of individual Cambridge School-style historical inquiry, requiring the assimilation of a massive number of Spanish, English, French, and Portuguese texts. If we extended our investigation of the spread of Marxism into the 20th century, for example, this problem would reach an astronomical level of complexity, as the USSR's ideological state apparatuses operated across 130 state-recognized domestic languages (Comrie 1981, p. 1) and over 60 foreign languages into which Soviet propaganda was translated and exported (Jacobs 2021).

The situation is not hopeless, however. By way of foreshadowing the possibilities enabled by modern computational-linguistic methods, for example, Google's Language-Agnostic BERT Sentence Embeddings (LaBSE) model (Feng *et al.* 2022) is trained to encode semantic relationships among words and sentences across 109 different languages, and is able to reason effectively across 35 additional languages due to "leaking" of semantic information from these languages into the texts used to derive the 109-language model[7]. Even more recently (and indicative of the rapid progress in the field over just two years), Facebook Research's "No Language Left Behind" model (Heffernan *et al.* 2022) nearly doubles this number, allowing researchers to analyze texts across over 200 languages via a single multilingual semantic space[8]. If, as we argue, researchers can use models like these to derive meaningful insights about the way words and phrases are deployed in political speech acts, in spite of their lack of fluency in the dozens of languages involved, it becomes possible to perform Cambridge School-style contextual studies on a massive scale—like that of Soviet political discourse—while allowing researchers to redirect their energies away from exhaustive close reading and language learning and towards the actual higher-level goal of the historical profession, namely, creatively synthesizing the available historical data to deepen our understanding of political thought.

In the remainder of this chapter, then, with this ultimate goal in mind, we build towards an understanding of the computational tools involved. Eschewing the gory details in favor of just that minimal set of information which a historical practitioner would need to know to perform

[7]See the "Support to Unsupported Languages" section of Yang and Feng 2020 for details on these 35 languages and what it means for this linguistic information to "leak" into the model.

[8]While Heffernan *et al.* 2022 provides details on how this model was trained, the model itself is available via Facebook Research's GitHub page.

their own computationally-aided inquiry, we emphasize throughout how the computational models implement and/or extend established principles of empirical social science which historians and political theorists already employ in their day-to-day practice. In addition, to concretely illustrate these connections to existing practices, we pair each newly-introduced tool or principle with examples of how they have been (or could be) fruitfully applied to existing problems from the historical and political-theoretic literature.

In the next section we introduce linguistic embedding algorithms, the central computational tool used in the remainder of the paper, describing how they explicitly combine information on *words* with information on their *contexts* to produce context-sensitive models of language. We then delve into the hermeneutic theories proposed by practitioners of the Cambridge School of intellectual history, emphasizing throughout how these context-sensitive language models are natural quantitative instantiations of the "general social and intellectual matrix" which Cambridge School practitioners claim one must analyze in order to understand the *meaning* of a text. We conclude our argument for the relevance of these computational language models in Section **??**, where we trace out the history of linguistic philosophy before and after the so-called "Linguistic Turn" of the 1950s, demonstrating how modern computational language models and Cambridge School studies in fact share a "common ancestor" in 1950s linguistic philosophy, when progenitors of computational linguistics like Noam Chomsky and key Cambridge School influences like J. L. Austin collaborated closely, only separating into "computational" and "humanistic" branches from the 1960s onwards.

We then turn from these motivating arguments to the main contributions of the paper: after discussing how basic embedding algorithms can be extended to incorporate models of *authorship* and *influence over time* in Section 3, we survey existing theories of influence in the history of political thought in Section 4.1, then present our central case studies in Section 4.2. In Section 4.3 we describe our study of "Text-Mining Influence Claims", then our study of the History of Marxism in Section 4.4, and finally our study of the communiques of the Intifada in Section **??**. We conclude in Section 5 with a recounting of the paper's contributions, followed by a model for future work which unifies our methodological contributions in the form of *embedding networks* through which researchers can precisely estimate the degree to which agents within a given social and intellectual

matrix (e.g., authors, newspapers, or institutions) influence one another over time.

# 2 Background

## 2.1 Word Embeddings: The Geometry of Political Thought

Although computational algorithms for analyzing text have existed since the 1950s, until recently the vast majority operated under a framework of drawing inferences from *word counts* within and across documents drawn from a corpus. Indeed, though we will eventually argue that these word-count-based approaches are insufficient for capturing semantic meaning within political texts—since they fail to capture the *contexts* in which the words are employed—they can nonetheless help researchers make basic inferences about the contents of a text corpus before embarking on a full context-sensitive study. For example, when faced with a new and unfamiliar corpus, a word-count-based approach could be helpful for quickly dividing the corpus into sub-categories based on pre-specified keywords.

As an illustration, we scraped plaintext versions of all New York Times articles from May 2022, including metadata on what section the articles were published in (Sports, Business, Arts, World News, etc.), and developed the intuitive schema for word-count based sorting shown in Table 1.

| **Section** | **Keywords** |
|---|---|
| U.S. News | `state`, `court`, `federal`, `republican` |
| World News | `government`, `country`, `officials`, `minister` |
| Arts | `music`, `show`, `art`, `dance` |
| Sports | `game`, `league`, `team`, `coach` |
| Real Estate | `home`, `bedrooms`, `bathrooms`, `building` |

Table 1: An intuitive schema for guessing what section of the New York Times a particular article was published in, using only the text of the article (i.e., by counting the number of occurrences of each keyword in each article)

With this schema in hand, we assigned each article a score for each potential section based on how many times that section's hypothesized keywords occurred in the article. By categorizing each article based on the section with the highest keyword counts (so that, for example, if the Food keywords appeared 5 times, the Technology keywords twice, and none of the other keywords

appeared in the article, it would be categorized as a Food article), these guessed categories match the articles' true NY Times sections an impressive 80.21% of the time, a significant improvement upon the 20% accuracy one would get (on average) via random guessing. Looking at the breakdown of this accuracy by section in Table 2, however, we can begin to see the fundamental limitations of this approach.

| | **Arts** | **Real Estate** | **Sports** | **U.S. News** | **World News** | **Total** |
|---|---|---|---|---|---|---|
| **Correct** | 3020 | 690 | 4860 | 1330 | 1730 | 11630 |
| **Incorrect** | 750 | 60 | 370 | 1100 | 590 | 2870 |
| **Accuracy** | 0.801 | 0.920 | 0.929 | 0.547 | 0.746 | 0.802 |

Table 2: The per-section accuracy scores for our word-count-based division of the *New York Times* article corpus.

While the Arts, Real Estate, and Sports sections use a vocabulary that is fairly unique to those sections (thus, for example, the word "bedrooms" almost never appears outside of the Real Estate section), distinguishing between U.S. News and World News is significantly more difficult, as many of the words we could try to use to distinguish between them still occur frequently in both sections. Knowing that the word "president" occurs frequently within an article, for example, without knowing the surrounding *context* in which it is used, does not help us very much in distinguishing between U.S. News and World News articles. And yet, crucially, if we had a method for *jointly* capturing the number of occurrences of the word *and* a breakdown of these occurrences by their contexts, suddenly our ability to distinguish between these two categories would skyrocket. For example, if we knew the following about two articles $A$ and $B$:

- The word "President" appears 10 times in Article $A$, with 5 of these appearances immediately preceding "Putin" and 5 immediately preceding "Clinton", and
- The word "President" appears 10 times in Article $B$, with all 10 of these appearances occurring immediately before "Clinton",

we could now easily distinguish that Article $A$ is likely to be a World News article while Article $B$ is likely to be a U.S. News article. With this insight, therefore—that purely word-count-based

approaches are insufficient for inferring meaning from text, but that the ability to capture the *contexts* in which words appear brings us much closer to this goal—we turn our attention away from these word-count-based models and towards a newer generation of explicitly *contextual* language models.

Specifically, in the remainder of this section we introduce *semantic embedding algorithms*, which enrich the analytical possibilities for text analysis by explicitly modeling the *context* of a given lexical unit (word, phrase, sentence, etc.) alongside the lexical unit itself. By incorporating this contextual information these algorithms can, as we'll see below, learn the *meanings* of words and phrases in a manner similar to that of human language acquisition—a manner summarized in the so-called Firthian Hypothesis: "You shall know a word by the company it keeps." (Firth 1957, p. 11) First, however, to illustrate the relevance of these considerations for the practice of political theory, we begin in the next section by outlining the intellectual-historiographic debate about what role *contextual knowledge* ought to play in our understanding of political-theoretic texts, especially in the aftermath of Quentin Skinner's provocative 1969 article "Meaning and Understanding in the History of Ideas" (Skinner 1969), which Pocock refers to retrospectively as "*the* manifesto of an emerging method of interpreting the history of political thought" (Pocock 2009, p. 128; emphasis added), namely, the method which became known as the Cambridge School approach.

## 2.2 The Historiography of Political Thought

Drawing on the linguistic philosophy of J. L. Austin, W. V. O. Quine, and the late Wittgenstein[9], the structuralism of Ferdinand de Saussure, and the pragmatics of H. P. Grice[10], scholars of the Cambridge School shifted much of contemporary historiography away from the notion of "perennial questions" in political thought (Bevir 1994), and towards a conception of historical texts as

[9]In the methodological volume of his 3-volume *Visions of Politics*, for example, Skinner explains that "Among philosophers of language, my approach owes most to" Wittgenstein's *Philosophical Investigations* (Wittgenstein 1953) and Austin's *How to Do Things with Words* (Austin 1962), along with Gottlob Frege's philosophy of language as summarized in Dummett 1973 (Skinner 2012, 161). The importance of Quine is asserted earlier in the volume: "when I read [...] that the holism espoused by Quine and Wittgenstein 'has had little impact on the philosophy of history', I feel that I have lived in vain", to which he adds that his "[Cambridge School] colleagues such as James Tully must feel the same." (Skinner 2012, p. 5).

[10]In the same methodological volume of *Visions of Politics*, Skinner describes his approach as "extending J. L. Austin's concept of a convention and [...] relating H. P. Grice's theory of meaning to Austin's account of illocutionary acts." (Skinner 2012, 133)

*interventions* into a particular, localized discourse.

Roughly speaking, an earlier school of intellectual historians (nowadays often associated with Leo Strauss) viewed the "great minds" of history—Plato and Aristotle, Machiavelli and Hobbes, Kant and Hegel—as engaged in a collective conversation on the "eternal questions" of philosophy, such as the question of what constitutes a good life or a good society. As summarized by Skinner, this school considered "a canon of leading texts" as "the only proper object of research in the history of political thought", given their unique contributions to "a set of perennial questions definitive of political thinking itself." (Skinner 1998, pp. 101–102).

Importantly—taking Strauss' view as representative of this trend—it is only this echelon of great minds who qualify as true political philosophers, since their philosophical work is "animated by a moral impulse, the love of truth", rather than the parochial desire to e.g. win an argument or persuade members of the public to adopt their preferences (Strauss 1959). In fact, under this conception, political philosophers must divorce themselves entirely from such local or day-to-day political concerns, as "it is only when the Here and Now ceases to be the center of reference that a philosophic or scientific approach to politics can emerge". In practice, then, the historian of political thought obtains the desired understanding of a text by "reading it over and over again", as

> We can learn more about their arguments by weighing them over and over again than by extending our knowledge of the circumstances in which they wrote. (Plamenatz 1963, p. x)

A Cambridge School approach, on the other hand, rejects the notion that a text can be understood outside of the day-to-day political issues of a thinker's time, i.e., outside of the circumstances in which they wrote. The claim is stated in its most direct form in Skinner's "Meaning and Understanding in the History of Ideas":

> There simply are no perennial problems in philosophy: there are only individual answers to individual questions, and as many different questions as there are questioners. There is in consequence simply no hope of seeking the point of studying the history of ideas in the attempt to learn directly from the classic authors by focusing on their attempted

answers to supposedly timeless questions. (Skinner 1969, p. 50)[11]

For instance, to take an obvious example, the Cambridge School would view Machiavelli's *The Prince* not as uninterested philosophical reflections on just rule and the structure of a just society, but instead as aiming to *do* something, to accomplish some desired end—in this case, to convince the new Medici regime to employ Machiavelli as a political advisor, after he had lost his patronage due to the fall of the previous regime. Or, to take a slightly more controversial example[12], Ellen Meiksins Wood's *Citizens to Lords* Wood 2008 challenges Strauss' conception of Plato's *Republic* as an uninterested, non-partisan tract on perennial questions of the good society, arguing instead that it cannot be understood outside of its sociopolitical context. For example, in her discussion of Plato's later years, she emphasizes the political aspirations behind the 385 BC establishment of his Academy. Serving as a kind of Hellenistic proto-think tank, "the political purposes of the Academy [were] unmistakable. Its students—the sons of wealthy Athenians and foreign families—were educated in Platonic politics and sent forth as consultants to rulers and cities throughout the Mediterranean world" (p. 66). Therefore, just as e.g. the content of George Mason University's economics seminars held for US federal judges cannot be understood outside of the political project of influencing legal outcomes (Ash *et al.* 2017), the content of Plato's writings cannot be understood outside of the project of trying to shape the political future of the Mediterranean world.

Taking inspiration from the insights into Plato's thought engendered by Wood's contextual analysis, we attempt to do the same in this work by contextualizing Marx's thought, in both its intellectual and material environment. Unlike in the case of Plato, however, where we only have access to the thought of a few of his key interlocutors[13], by Marx's time literacy, communication technology, and intellectual institutions were widespread enough to make a deep reading of all his

[11]The statement appears in slightly modified form in Skinner 2012, p. 88: "there are no perennial problems in philosophy. There are only individual answers to individual questions, and potentially as many different questions as there are questioners.".

[12]"Controversial" in the sense that practitioners of the Cambridge School approach, along with Wood herself, view the approach of this work as differing in some respects from the "orthodox" Cambridge School approach established in the works of e.g. Skinner, Pocock, and John Dunn (specifically, in its attempt to infuse the "purely" discursive or illocutionary analyses of e.g. Skinner with insights drawn from historical materialism).

[13]Importantly, we emphasize that this was not because he only *had* a small number of interlocutors, just that we only have *access* to a small number of them due to the selective preservation and transmission of historical texts (Shapiro *et al.* 1987), an effect compounded by the lower proclivity to record systematic thought in the form of writing in Plato's time, relative to Marx's.

interlocutors impossible, thus necessitating the computational "distant reading" methods detailed in the previous section.

In fact, we emphasize the links between our study and those of the Cambridge School because we view the methods developed herein simply as ways to overcome an inherent limitation of the latter's research program as it has been conducted thus far. The genre-defining Cambridge School studies—Skinner's *Foundations of Modern Political Thought*, Pocock's *Machiavellian Moment*, and John Dunn's *The Political Thought of John Locke*—all analyze political thinkers from *before* the era of mass literacy and rapid global communication networks, the era that Marx was born into. The subjects of these studies typically wrote with a small audience—a set of key political figures they hoped to influence—in mind. The first volume of Skinner's *Foundations of Modern Political Thought*, for example, traces the development of the mirrors-for-princes genre across the independent city-states of Italy in the thirteenth century, emphasizing how "the ambition to supply their rulers and magistrates with practical advice on how best to conduct themselves provides the central theme" for these works. Given this proscribed discursive community, restricted essentially to the authors alongside the few prominent rulers and magistrates of their polities for whom they wrote, Skinner is able to track incremental changes in the genre over time by examining how political issues in these city-states (for example, the power struggles between the Empire, the Pope, and their own desire for independence) gave rise to the novel ideologies proffered in these texts. But, as Cambridge School works move towards examining later, post-Renaissance communities of discourse, e.g. that of the English radicals to whom Locke wrote from his exile in Holland[14]), they are increasingly hampered by the need to study of larger and larger—and thus increasingly intractable—collections of texts, if they hope to gain a representative understanding of the intellectual context of key texts like Locke's.

So, while the limited scope of the most impactful Cambridge School works is what allows their tractable contextual analyses via deep reading, it simultaneously erects an artificial barrier whereby texts written for mass audiences are excluded from this mode of analysis, an exclusion with no political-theoretic or historiographic justification. Thinkers like Kant, Hegel, and Adam Smith, for

[14]See, e.g., Ashcraft 1986.

example, who shaped political thought in the midst of the industrial revolution's earliest upheavals, therefore form a sort of "blind spot" in Cambridge School analyses: should their works be analyzed as discussions restricted to a small set of political and intellectual elites (so that understanding can be achieved via deep reading and comparison of texts with the intrigues of these political and intellectual elite), or do they represent the point at which historians need to start taking into account the burgeoning mass publics of the subsequent centuries? We hope, therefore, to point the way towards an expanded Cambridge School methodology which employs both deep reading *and* computational inference to eliminate this artificial time-barrier, thus allowing researcher to employ the power of the Cambridge School approach towards understanding the *full* development of political thought, in uninterrupted sequence, from antiquity up to the present.

Looking towards our studies in the following chapters, for example, a study of the development and trajectory of Marx's thought cannot plausibly restrict itself to studying the discursive communities of one country, one ideological tendency, one language, or one "epoch" of 19th century history. The key elements of his critique of capitalism were forged over the course of a decade—the 1840s—in which a whirlwind of revolutionary upheavals and governmental expulsions swept him from Germany to France, Belgium, and England, where he finally settled after the abortive revolutions of 1848. Historians thus generally agree that his thought represents a mixture of German philosophy, French socialism, and British political economy, but disagreements begin to arise when the details of this mixture are interrogated: What particular concepts did he absorb from each, and to what extent did he modify or transform them? Did the influence occur gradually, through e.g. his day-to-day interactions with workers in Paris? Or can we pinpoint particular moments when his reading of certain texts immediately affected his thought? And, while we know that analyses at this level of granularity have been carried out to great effect for earlier time periods—as we saw above for the case of Skinner's study of the step-by-step progression of the mirrors-for-princes genre—is this type of detailed understanding still possible when the discourse is carried out across so many countries, languages, and intellectual communities?

In the eyes of Cambridge School practitioners themselves, this indeed represents uncharted territory. In a 2009 collection of reflective essays Pocock acknowledges these linguistic limits,

conceding that "most of the work produced in Cambridge and by those associated with it has been concerned with [...] a history of political thought largely anglophone." (Pocock 2009, p. 140) He then draws attention to the temporal limits as well, pointing out that "Skinner has not yet ventured far into the eighteenth century", and adding, "I myself cease from inquiry about the year 1790" (*ibid.*, p. 141). After discussing some promising historical studies of the post-1790 development of political thought, he nonetheless concludes that "A Skinnerian approach to the modern and the post-modern has not yet been tried." (*ibid.*) This is precisely the project we will embark upon in later chapters, but it is a project for which we'll need additional hermeneutic machinery if we want to manage the increased complexity of nineteenth- and twentieth-century discourse, relative to the pre-1790 "neo-Latin culture in which discourse was the preserve of established clerisies operating stable and continuous languages." (*ibid.*) Indeed, Pocock's 1790 cutoff is particularly noteworthy given our motivating interest, our aim of understanding how exactly the emergence of mass publics affected the conduct of political theorization and argument. As Carla Hesse summarizes the effect of the French Revolution on the publishing industry, for example,

> "Between 1789 and 1793, [the Revolution] swept away the monopolies that France's preeminent cultural elites had over the means of producing and disseminating ideas through the printed word" (Darnton *et al.* 1989, p. 82)

Pocock's and Skinner's decisions to restrict their close-reading-based studies to the era before this radical shift in European print culture are thus understandable: while the sweeping changes described by Hesse already portended a massive expansion of discursive communities *within* France[15], the subsequent spread of French norms and institutions via Napoleon's incursions transformed this national trend into a continent-wide phenomenon by 1815, inaugurating "a modern age in which the printing press, polite conversation, and public opinion were the defining discursive institutions." (Goodman 1996, p. 5)

Thus, before beginning our post-1790 Cambridge School investigation of this "modern age", we argue in the next section that computational-linguistic methods developed in recent years enable

[15]Later in the essay Hesse describes one immediate effect of this "unprecedented expansion and democratization" of discourse in the proliferation of political, intellectual, and scientific journals: "The number of journals produced in Paris skyrocketed from 4 in 1788 to 184 in 1789, and 335 in 1790." (Darnton *et al.* 1989, p. 92)

us for the first time to tackle this increased complexity which "progressively aris[es] as clerisies are replaced by intelligentsias and political discourse becomes increasingly demotic" (Pocock 2009, p. 141). By identifying connections between the principles of Cambridge School hermeneutics and the inner-workings of these computational tools in Section 2.3, we derive a neo-Cambridge School methodology for developing verifiable, reproducible answers to Cambridge School-motivated context-sensitive questions—given a set of assumptions regarding how their terms should be operationalized—that enables us to bring this approach to bear on developments in nineteenth- and twentieth-century political thought.

Importantly, our phrasing here (the "given a set of assumptions" qualifier) is intended to highlight a key difference between our approach and the approach of what we'll call "vulgar" computational social science[16], namely, that the answers we generate are fundamentally *conditional* on the set of assumptions that have been encoded in the form of the algorithms' initial conditions and parameters. In other words, while we do attempt to justify the prior assumptions we encode in the algorithms throughout, the key contributions lie not in these assumptions or their justifications (which we hope can simply be seen as provisional and contestable "encodings" of already-existing arguments from the literature) but rather in the veracity of our methods with respect to their ability to "scaffold" *all* sides of a given interpretive debate with verifiable and reproducible metrics. This endeavor can be considered successful, for example, to the extent that it allows historians of political thought to adopt a procedure akin to Rawls' "reflective equilibrium" (Rawls 1951), where evidence generated via our methods can be used to increase or decrease one's confidence in a given interpretation of a historical text or thinker.

## 2.3 From Computational Linguistics to the Cambridge School and Back

Our application of computational-linguistic methods to questions of intellectual history in fact has a rich history, one which was unfortunately lost in the institutional divisions-of-labor which followed. Our aim, in fact, is simply the *re*-integration of both the Cambridge School historiographic approach and modern contextual computational-linguistic methods with their "common ancestor", a strand

[16]This terminology is meant simply to denote a model of the historiographic tendencies we will aim to avoid, not to suggest that any particular researcher or work actually adopts this "vulgar" approach.

of thought catalyzed by the rapid turn away from logical positivism and towards the analysis of natural language in the field of linguistic philosophy over the course of the 1950s.

Rooted in the late Wittgenstein's *Philosophical Investigations* (Wittgenstein 1953)[17] and culminating in J. L. Austin's *How to Do Things with Words* (Austin 1962), this approach broke sharply with the theretofore dominant logical-positivist strand of linguistics. Philosophers like Austin argued that statements such as "I hereby declare you husband and wife" or "Look out!", despite not being translatable into logical true/false statements (and thus rejected outright as "meaningless" by logical positivists[18]), ought to be analyzed by linguists nonetheless in terms of their "locutionary effects"—what their utterances "do" in the world when spoken—rather than their internal structure or consistency with other logical statements.

In the whirlwind of philosophical activity sparked by the insights in Austin's work, we argue, the work of linguists interested in computation and of those interested in literary or cultural aspects of language became "decoupled". Indeed, one can see in the work of e.g. Roman Jakobson the lack of separation between these two endeavors: simultaneously a professor of Linguistics and Czech Literature, his 1948 book on the 12th-century Russian epic *The Tale of Igor's Campaign* (Grégoire *et al.* 1948) and his 1960 essay "Linguistics and Poetics" (Jakobson 1960) stand alongside highly-technical works like *Child Language, Aphasia and Phonological Universals* (Jakobson 1941) and his 1957 MIT lecture on "Linguistics and Physics" (Jakobson 1957).

The pathway leading from the logical-positivism-counterposed work of J. L. Austin, Jakobson, and others to modern contextual computational linguistics is easy to trace: Noam Chomsky, for example, explicitly viewed his own work as a "continuation" of Jakobson's technical work, and his 1955 PhD dissertation established a framework for understanding which linguistic structures were and were not capable of unambiguous parsing by a computer[19]. That same year, after attending the

[17]Though the genesis of this "turn" is indeed neatly traced back to the late Wittgenstein in most popular narratives, precedents for Wittgenstein's contributions did certainly exist in the works of less-often-acknowledged linguists and linguistic philosophers such as Roman Jakobson (who was, in turn, strongly influenced by Ferdinand de Saussure's 1916 *Course in General Linguistics* (Saussure 1916), as was the Cambridge School's J. G. A. Pocock).

[18]Our narrative here, for the sake of brevity, papers over the subtleties of the interplay between logical-positivist and natural-language philosophers. The oft-cited final proposition of the early Wittgenstein's 1921 *Tractatus Logico-Philosophicus* (Wittgenstein 1921), however, that "whereof one cannot speak, thereof one must remain silent," broadly summarizes the perspective of early logical positivists with respect to non-logically-analyzable linguistic utterances.

[19]Regular expressions, for example, allow more powerful searching within and across web pages, as these pages are scanned not for specific letters or digits but for patterns corresponding to "regular" context-free grammars, one of

Harvard lectures published as *How to Do Things with Words*, Chomsky met and began regularly corresponding with Austin, who in turn discussed Chomsky's dissertation in a series of meetings in 1957.

The diverging pathway leading from Austin and Jakobson to e.g. Quentin Skinner is also straightforward, as the latter's 1968 "Meaning and Understanding in the History of Ideas"—considered the canonical reference for the central tenets of the Cambridge School—heavily cites Austin's works, and Skinner explicitly mentions Austin as a (perhaps the) key influence in interviews conducted in the years since its publication. For the purposes of our argument, however, we focus on a 1957 publication situated directly at the intersection of these two strands, J. R. Firth's "A Synopsis of Linguistic Theory, 1930–1955". In this work, essentially providing a motto and *raison d'etre* for both the Cambridge School and for contextual computational linguistics, Firth made his famous assertion quoted above, that "you shall know a word by the company it keeps".

Following the pathway from Firth and Austin to subsequent linguistic philosophy, we can interpret W. V. O. Quine's "*gavagai*" thought experiment[20] published three years later in his 1960 *Word and Object* (Quine 1960) as driving Firth's point home by illustrating the indispensability of context for the construction of linguistic meaning and understanding. Quine begins by asking the reader to imagine themselves trying to learn a foreign language by observation. Upon seeing a rabbit run by, they observe the native speakers of the language pointing and saying "gavagai". Though our instinct may be to infer thereby that "gavagai" must mean "rabbit", in fact with only this information to go by it may just as well mean "furry thing", "running animal", or myriad other possibilities. It is only by observing multiple different contexts in which "gavagai" is and is not used to denote various objects or events that one can essentially narrow in on, or "triangulate", the meaning of "gavagai".

Contextual embeddings, the primary tool used throughout our study to measure and compare the linguistic meanings of words, phrases, sentences, and documents, are constructed precisely to "quantify" this Firthian notion of meaning: words[21] are represented by points in a geometric space

the four levels of grammatical complexity within the Chomsky Hierarchy.

[20] In recent computational linguistic work, the word "wampimuk" is often used in place of "gavagai", following Lazaridou *et al.* 2014.

[21] We focus on word embeddings for simplicity in this section, but note that the principle generalizes to phrases,

such that the distance between these points is proportional to the likelihood that they appear in similar contexts in a corpus. Under the hood, this entails training a neural network to predict, given a particular word randomly sampled from a sentence, what the surrounding words in the sentence are. Thus, for example, a neural net which predicts "Scotia" when given the word "Nova" will likely do better on this task than one which predicts "pineapple", ceteris paribus. The information which the trained neural network has stored for each word is then used to construct the geometric space: since the optimized network placed a high probability on seeing "Scotia" in the context of the word "Nova", the distance between the geometric points corresponding to "Nova" and "Scotia" is smaller than that between the points corresponding to "Nova" and "pineapple".

Drawing the parallel between this and the *New York Times* corpus discussed above, for example, upon seeing the word "Clinton" the neural net would probably achieve a high accuracy score by predicting that "president" will appear somewhere in this context. Since the same is true of seeing the word "Putin", therefore, the resulting geometric space will map "Clinton" and "Putin" to points very close together, representing the fact that they are used in similar contexts. Similarly, the points for "Clinton" and "U.S." will be close together, as will the points for "Putin" and "Russia".

The multiple types of relations thus captured in the geometric space (since the simple fact that "$X$ often appears in context $Y$" is still purposefully ambiguous with respect to the precise *nature* of the relationship between $X$ and $Y$) will in fact give rise to higher-order geometric relationships between points than just "close together" or "far apart". The aforementioned points, for example, will be arranged such that one can perform "analogical math" on the terms:

$$\overrightarrow{\texttt{Putin}} - \overrightarrow{\texttt{Russia}} + \overrightarrow{\texttt{U.S.}} = \overrightarrow{\texttt{Bush}}, \tag{1}$$

$$\overrightarrow{\texttt{Washington}} - \overrightarrow{\texttt{Bush}} + \overrightarrow{\texttt{Moscow}} = \overrightarrow{\texttt{Putin}}, \tag{2}$$

$$\overrightarrow{\texttt{U.S.}} - \overrightarrow{\texttt{Iraq}} + \overrightarrow{\texttt{Russia}} = \overrightarrow{\texttt{Ukraine}}, \tag{3}$$

and so on. Thus we can use these *entity* vectors to solve for vectors representing *relationships* between these entities—a technique that will become central to our studies in future chapters, for

sentences, and documents, as we discuss in more detail in Section 3 below.

example, when we analyze the different relationships Marx and Hegel ascribe to the (`state`, `civil society`) entity pair[22]:

$$\overrightarrow{\textsf{Leader-Of}} = \overrightarrow{\texttt{Putin}} - \overrightarrow{\texttt{Russia}}, \tag{4}$$

$$\overrightarrow{\textsf{Invaded}} = \overrightarrow{\texttt{U.S.}} - \overrightarrow{\texttt{Iraq}}, \tag{5}$$

$$\overrightarrow{\textsf{Capital-Of}} = \overrightarrow{\texttt{Ukraine}} - \overrightarrow{\texttt{Kyiv}}. \tag{6}$$

This means that, despite not having this knowledge explicitly encoded into the system beforehand, the algorithms can infer these relationships from the fact that words for capitals like "Washington" and "Tehran" are often used in similar contexts, metonymically, to refer to states and/or administrations. In just the corpus of January–June 2006 *New York Times* articles, for example, the following sentence alone provides crucial information to the algorithm regarding the country-capital relationships Capital-Of(`Tehran`, `Iran`) and Capital-Of(`United States`, `Washington`), as well as the existence of diplomatic interactions between the two states:

> "The `United States` has said that if `Iran` suspended uranium enrichment, `Washington` would join the Europeans in direct talks with `Tehran`."

In fact, even in corpora without individual sentences like these directly establishing the relationships among entities, their relationships can be inferred via the *pooling* of information across multiple sentences. By synthesizing information from sentences like the following (also from the 2007 *New York Times* corpus), for example, the same relationships can be inferred:

(a) "A `British` official, speaking anonymously because of the delicacy of the diplomatic exchanges between `London` and `Tehran`..."

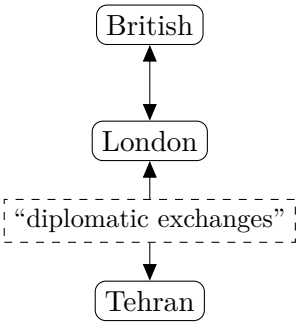
(b) "The `Bush` administration said Thursday that the release of 15 `British` sailors and marines held by `Iran` for two weeks created no new openings in dealing with `Tehran`..."

(c) "By deceiving the nuclear agency about its activities, President `Bush` and `British`, French and German officials say, `Iran` has given up whatever treaty rights it once enjoyed."

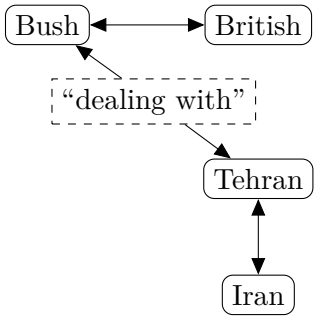
[22] We denote these *relational* vectors using a sans-serif font ($\overrightarrow{\textsf{Relationship}}$) to differentiate them from the entity vectors, which are instead rendered in a fixed-width font ($\overrightarrow{\texttt{Entity}}$).

(d) "Asserting `Washington`'s determination to protect Japan and South Korea, its principal allies in the region, Mr. `Bush` said the `United States` 'will meet the full range of our deterrent and security commitments.'"

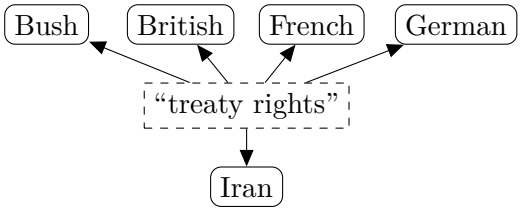
Thus, considering these sentences in order, the algorithm could construct networks of meaning for each sentence as illustrated in Figure 1, then combine the information from these individual networks as shown in Figure 2, resulting in the synthesized-information network shown in Figure 3, where the capital cities for the UK, US, and Iran are situated similarly with respect to the `Bush` node.



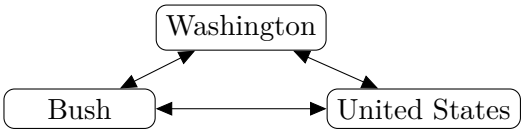

(a) "A `British` official, speaking anonymously because of the delicacy of the diplomatic exchanges between `London` and `Tehran`..."




(b) "The `Bush` administration said Thursday that the release of 15 `British` sailors and marines held by `Iran` for two weeks created no new openings in dealing with `Tehran`..."



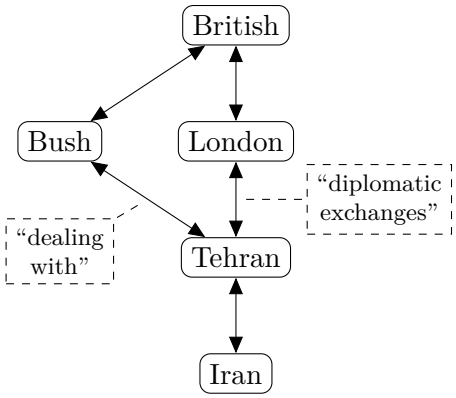

(c) "By deceiving the nuclear agency about its activities, President `Bush` and `British`, French and German officials say, `Iran` has given up whatever treaty rights it once enjoyed."



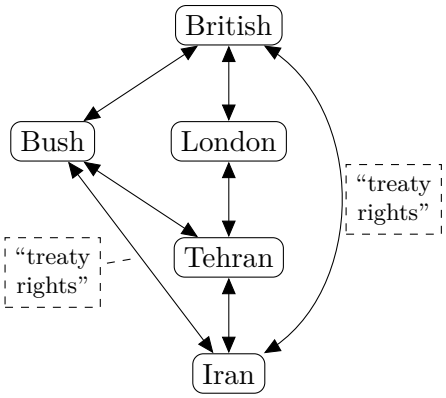

(d) "Asserting `Washington`'s determination to protect Japan and South Korea, its principal allies in the region, Mr. `Bush` said the `United States` 'will meet the full range of our deterrent and security commitments.'"

Figure 1: Graphical models of the relationships which context-sensitive text-analysis algorithms can infer from sentences (1) through (4) above.

The meaningful spatial positions of the learned vectors, moreover, give rise to a type of transitive property allowing us to infer information that is *never* stated in the corpus, directly or indirectly. To see this, note how in Figure 4 one can travel from the $\overrightarrow{\texttt{Russia}}$ vector to the $\overrightarrow{\texttt{Kyiv}}$ vector via




(a) The systematic merging of the networks in Figures 1a and 1b.




(b) The network formed by incorporating the information in Figure 1c into the network of panel (a).

Figure 2: An illustration of how, after constructing the individual-sentence context networks in Figure 1, the algorithm can systematically merge the information across sentences, eventually producing the final corpus-wide network shown in Figure 3.

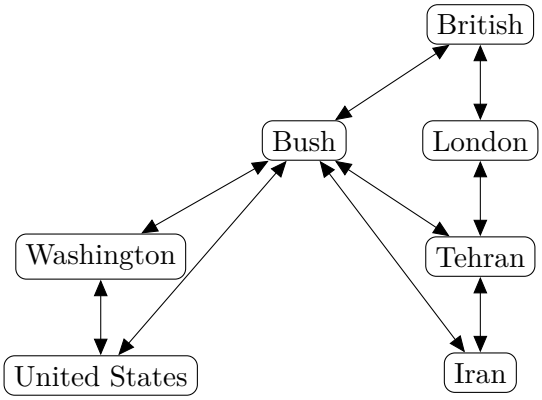



Figure 3: The final corpus-wide contextual network, combining all information from the individual networks in Figure 1 (edge labels omitted for clarity).

many different pathways, for example by:

(a) Following the $\overrightarrow{\texttt{Invaded}}$ vector to $\overrightarrow{\texttt{Ukraine}}$ and then following the $\overrightarrow{\texttt{Capital-Of}}$ vector to $\overrightarrow{\texttt{Kyiv}}$,

(b) following the $\overrightarrow{\texttt{Capital-Of}}$ vector to $\overrightarrow{\texttt{Moscow}}$ and then following the $\overrightarrow{\texttt{Invaded}}$ vector to $\overrightarrow{\texttt{Kyiv}}$, or

(c) Following the $\overrightarrow{\texttt{Leader-Of}}$ vector to $\overrightarrow{\texttt{Putin}}$, the $\overrightarrow{\texttt{Invaded}}$ vector to $\overrightarrow{\texttt{Zelensky}}$, the $\overrightarrow{\texttt{Leader-Of}}$ vector in the opposite direction to $\overrightarrow{\texttt{Ukraine}}$, and finally the $\overrightarrow{\texttt{Capital-Of}}$ vector to $\overrightarrow{\texttt{Kyiv}}$.

We illustrate this multitude of paths in Figure 5, where the pathways described by (a), (b), and

(c) above are colored and labeled accordingly.

For example, having learned the aforementioned $\overrightarrow{\text{Capital-Of}}$ vector based on the respective contexts in which "U.S." and "Washington" were used in the corpus, one could then add this vector to the vector for the entity `Poland` to obtain the vector for `Warsaw`, even if the corpus contained no explicit statements about Warsaw being the capital of Poland. Typically this is made possible by the algorithm's ability to pick up on subtle contextual patterns, like the fact that capital cities of a given country are often the most frequently-mentioned cities in stories about those countries. Thus, for example, the embedding for `Ulaanbaatar` could occupy a similar position relative to `Mongolia` as `Warsaw` does to `Poland`, even without the former appearing together in an explicit "capital of" context in the corpus, on account more simply of their common co-occurrence.

It is in their ability to capture meaningful semantic relationships like these—an ability extending even to relationships never explicitly mentioned within a corpus—that we argue for understanding word embedding algorithms as methods for inferring and studying the "geometry of political discourse".

While we will dive into the details of these embedding algorithms more deeply in Section 3 below, to see *how* they are able to capture these multi-layered semantic relationships, for now we conclude by re-emphasizing how understanding contextual embedding algorithms as *computational implementations* of Firth's hypothesis thus "completes the loop" from computational linguistics to the Cambridge School and back: the contextual analysis advocated by Cambridge School scholars and carried out via a deep reading of a wide array of historical texts is precisely mirrored in the processes that generate these neural embedding spaces. And, as touched on above (a consideration which we will also discuss in more detail below), while embeddings restricted to the level of words are perhaps too lexically fine-grained to allow a general understanding of a text in the sense advocated by the Cambridge School, the move from contextual word embeddings up to contextual *sentence* embeddings in recent years brings us significantly closer to the textual level of analysis, and offers hope that the types of methods used to aggregate information from the word to the sentence level could also bear fruit for the further aggregation of information up to the level of full texts within a corpus.

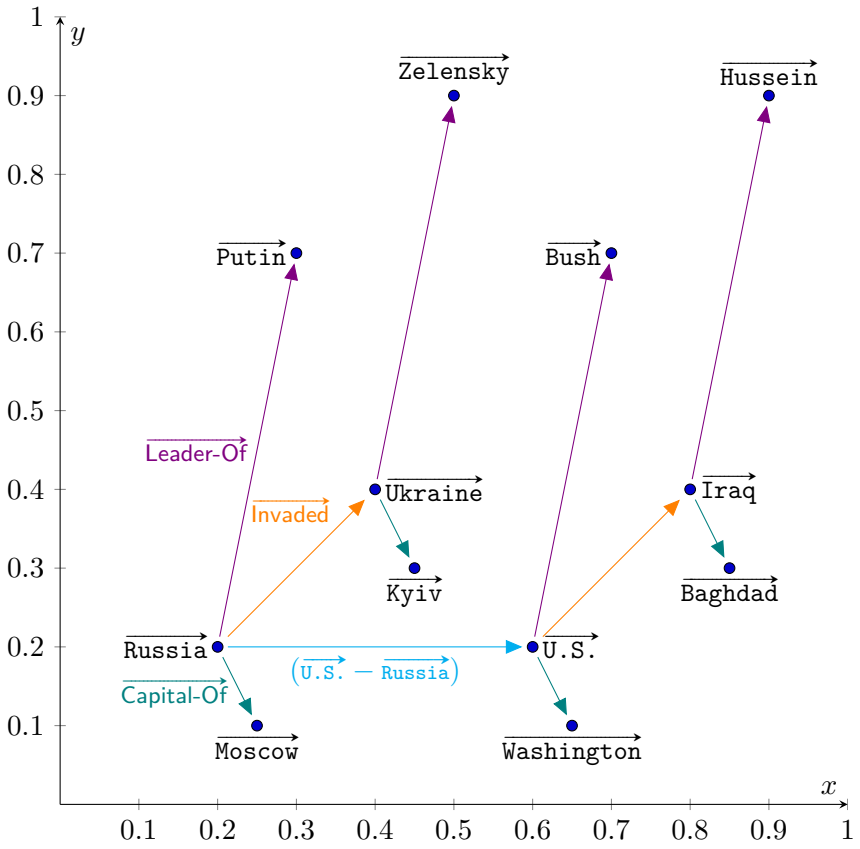



Figure 4: A visualization of the "analogical math" which can be performed within word embedding spaces, due to their ability to capture multiple *types* of word-word relationships in the form of distances ranging over multiple dimensions: entity vectors are typeset in fixed-width font (`Entity`), while relational vectors are typeset in serif font (Relation). The vector from `Russia` to `U.S.`, a relational vector, is typeset in fixed-width font only to illustrate its mathematical form (i.e., that it is computed via the subtraction two entity vectors).

# 3 Models of Meaning and Context

## 3.1 Constructing Contextual Fields

The key computational tool used throughout the studies in the remainder of this work is the contextual sentence embedding algorithm, a method for mapping linguistic units (words, phrases, sentences, documents) into geometric spaces based on their semantic content. In fact, there is no single "embedding algorithm", but rather a wide range of algorithms which transform various linguistic entities into points within a high-dimensional geometric space. Although our unit of analysis herein is a *document* (e.g., a text, pamphlet, or letter), these documents are in general

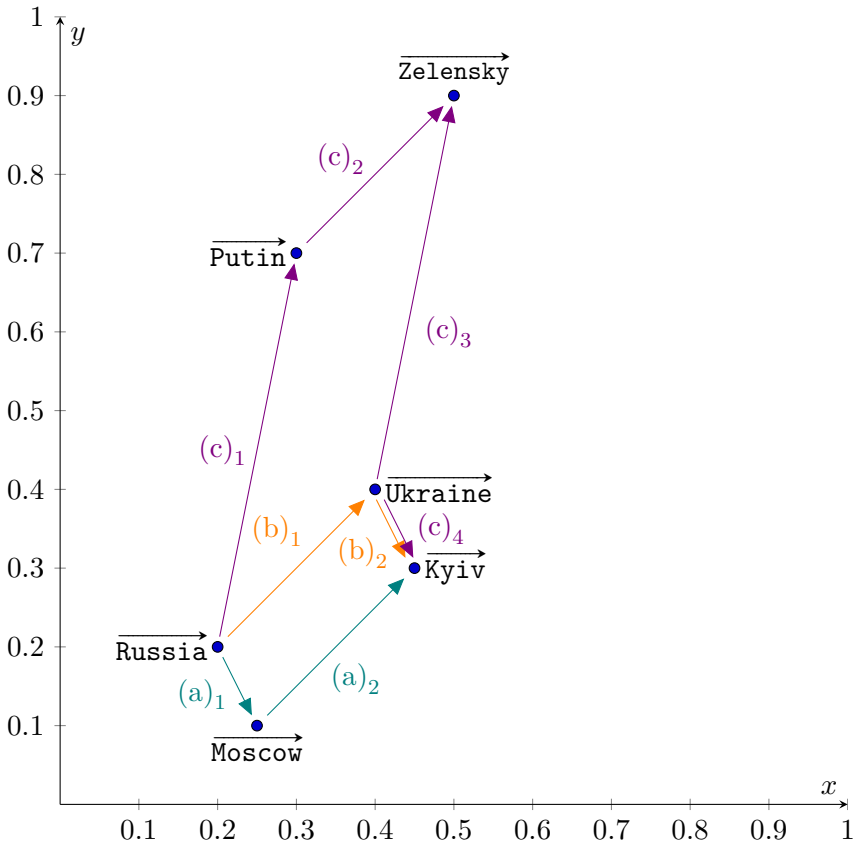



Figure 5: Visualizing three of the potential pathways by which one can move from the `Russia` to the `Kyiv` vectors.

too long for a single embedding algorithm to handle – state-of-the-art embedding algorithms like `BERT` have an upper limit of 512 tokens that can be jointly encoded into a single high-dimensional vector. Thus we instead compute a separate embedding for each *sentence* within a text using `SentenceTransformers` (Reimers and Gurevych 2019), then combine these sentence vectors into a single document vector via mean pooling. As a robustness check, however, we computed document-level embeddings via an experimental document embedding method called Longformer, described in Beltagy *et al.* 2020, and obtained qualitatively similar results.

Turning to the issue of *how* exactly the semantic information in a text is given a geometric interpretation: at the most basic level, sentence embedding algorithms take every word appearing in a corpus and represent them as points within a geometric space, such that words which are

used in similar contexts[23] will be placed closer together in the space than sentences which use dissimilar words and/or dissimilar contexts around these words. In this way, computers can begin to reason about semantic relationships between words by analyzing the geometric properties of these constructed spaces, as in Figure 6.

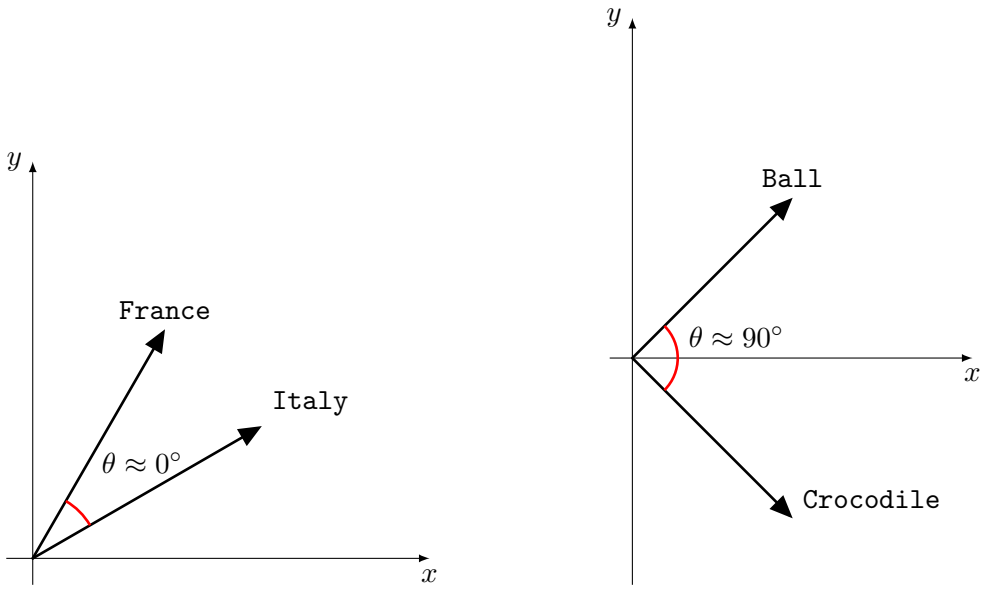



Figure 6: The word-level linguistic intuitions which, by constructing an appropriate geometric space, can be quantitatively measured by way of the Cosine Similarity metric: (a) On the left, `France` and `Italy` are quite similar, so that $\theta$ is close to 0°, and $\cos(\theta)$ is close to 1. (b) On the right, `Ball` and `Crocodile` are not similar, so that $\theta$ is close to 90°, and $\cos(\theta)$ is close to 0.

By implementing this micro-level property—that is, the property that words or phrases which share similar contexts should be close together in the generated geometric space—across an entire text corpus, what emerges at the macro level is a landscape of the words and phrases in the corpus within which distinct regions can be identified which correspond to our intuitions about semantic meaning in human language. For example, Figure 7 visualizes a cross-lingual embedding space where each node is colored based on what part of speech it represents, demonstrating how these embedding algorithms, despite focusing solely on micro-level (i.e., word- or phrase-level) similarity properties, naturally give rise to geometric spaces which capture macro-level properties of human

[23] Although "context" can be operationalized in different ways based on what information a user hopes to extract, in our case the context of a word $w$ in a sentence $S$ is defined to be the set of $n$ words appearing before and after $w$ in the sentence. For example, if $S$ is "`The sleepy grey cat likes salmon.`", and $w$ is "`cat`", then the context of $w$ with $n = 2$ would be the set {`sleepy`, `grey`, `likes`, `salmon`}.

languages.

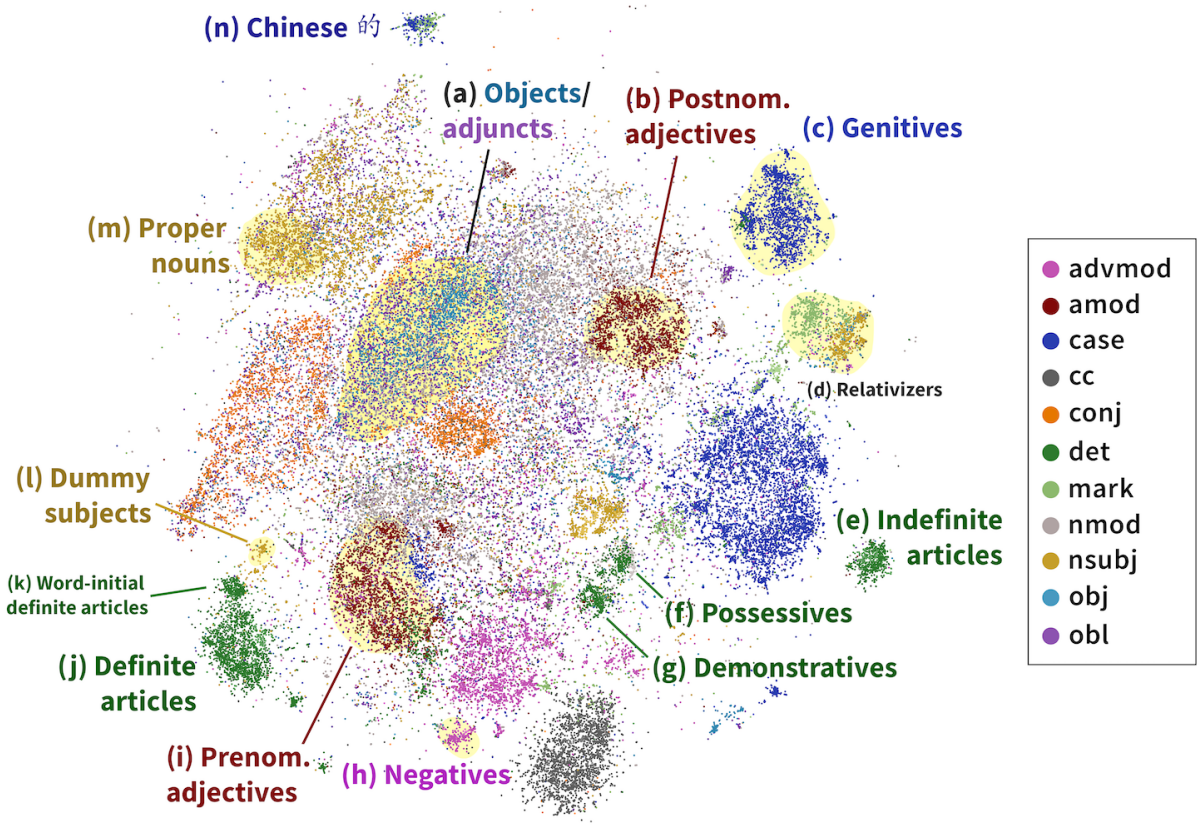



Figure 7: An example of the higher-level properties of language which naturally emerge when an embedding algorithm computes *word-level* contextual-semantic similarities (that is, moves the points for words which occur in similar contexts closer together) across a large corpus of text from many different languages. (From Chi *et al.* 2020, Figure 6)

To understand how exactly this text-to-geometric-space transformation works, Figure 9 illustrates how the task can be represented as a mathematical problem which, when solved, produces precisely the desired semantic-capturing and context-sensitive geometric space. Given the *full* set of contextual information about each word—that is, how many times the word appears in each possible context—the aim is to compress the information for each word into just 3 dimensions, such that words which appear in similar contexts will be close together when visualized in 3D space. Visualizing the 3D space generated in the figure, for example, word 5 will be very close to word $n$, as word 5 will be mapped to the coordinate $(x, y, z) = (.2, .0, .3)$, while word $n$ will be mapped to $(x, y, z) = (.2, .1, .3)$, so that they differ only by 0.1 in the $y$ coordinate, as visualized in Figure 10.

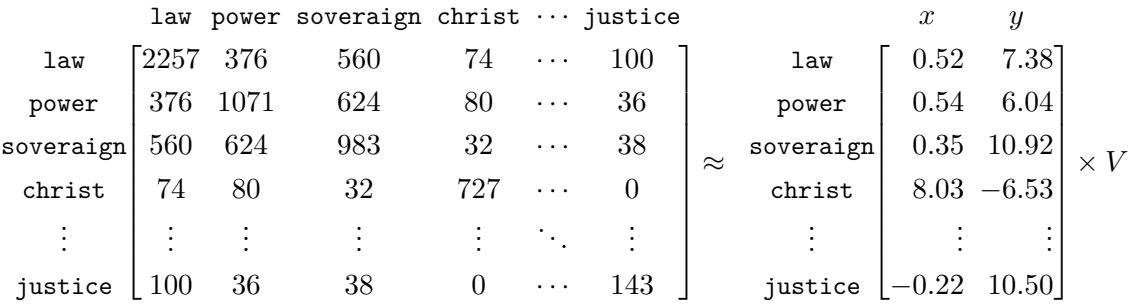



Figure 8: The mathematical operation (dimensionality reduction) by which we compress the *full* set of word-context information in Hobbes' *Leviathan* (the matrix on the left) into a smaller matrix of two-dimensional vectors (the rows in the right-side matrix) which best preserve contextual similarity. That is, similar rows in the left-side matrix will be mapped to close-by points in 2D space, with $(x, y)$ coordinates as given in the right-side matrix.

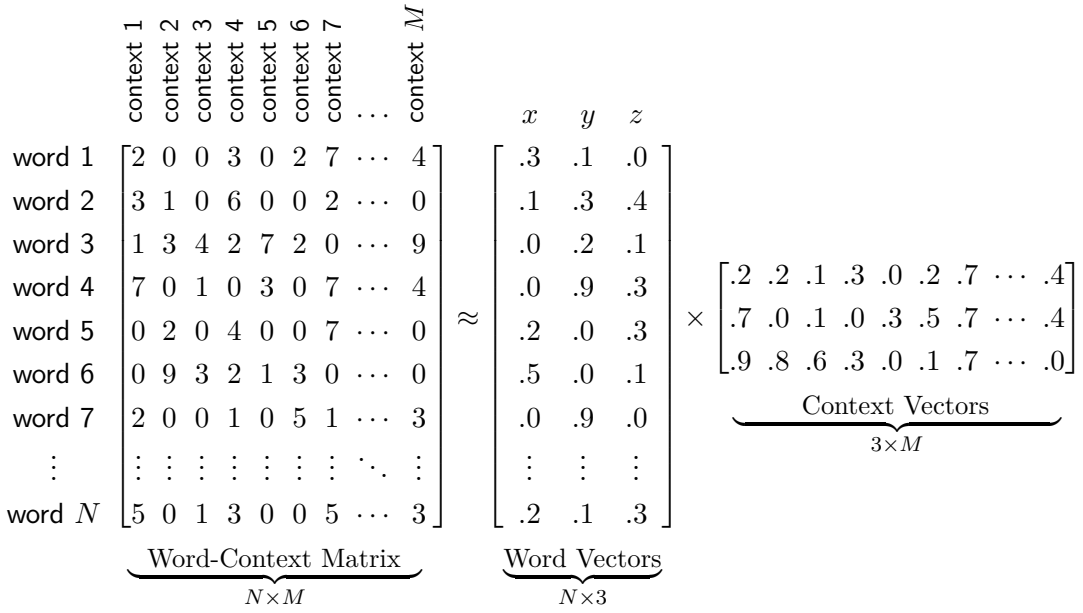



Figure 9: An example of the matrix decomposition procedure that word embedding algorithms implement, to solve the problem of *retaining* information about word-context relationships while *reducing* the $M$-dimensional representations of each word down to 3 dimensions. (Example adapted from Kozlowski *et al.* 2019, Figure 1)

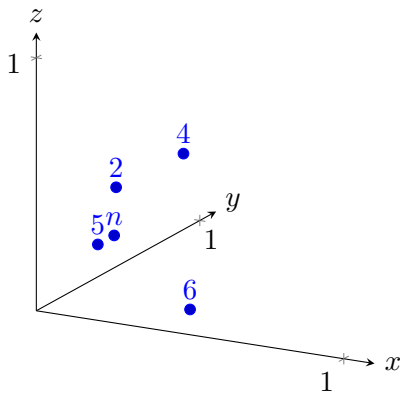



Figure 10: A plot of the 3-dimensional word representations in Figure 9.

To see how powerful this dimensionality-reduction approach is, let's consider an extremely small corpus of four "documents", where each document is actually just a single 3-4 word sentence:

1. `Green eggs are sweet.`
2. `Ham is sour.`
3. `Cats love sweet food.`
4. `Dogs love sour food.`

Taking the set of unique words[24], and defining the context of a word to be the full sentence in which it appears, we obtain the Word-Context Matrix shown (along with the resulting two-dimensional embeddings) in Figure 11. By examining the plot of the resulting two-dimensional embeddings in Figure 12, we can see that despite having a corpus of only four sentences (providing a total of 15 tokens) to infer information from, the method is already able to detect multiple semantic relationships: First, we see that it has captured the identical context-employment of `green` and `eggs`, by mapping them onto the same point in the 2D space. Second, we see that it has used the $x$-axis to separate two triangles of points, one representing the semantic relations involving `cat` and the other the semantic relations involving `dog`.

[24]That is, the unique words after the *function words* "are" and "is" have been removed. Standard practice in word embedding models is to remove all such function words, since they pertain to the *syntax* rather than the semantic content of a given sentence. In other research applications, however, the syntactic structure of the sentences may be exactly what one is hoping to capture with these embeddings, in which case all *non*-function words are removed. See Mosteller and Wallace 1964 for a seminal application of this syntax-based analysis in the social sciences.

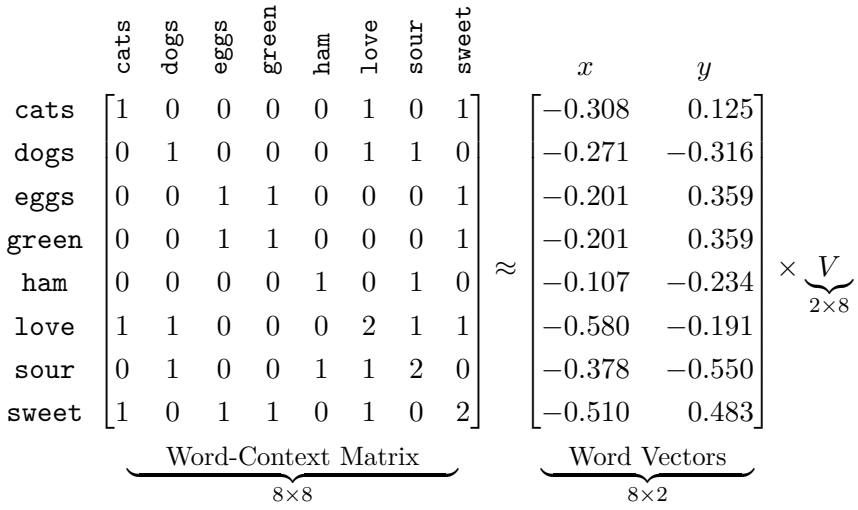



Figure 11: The Word-Context Matrix for the four-sentence Green Eggs and Ham corpus, along with the two-dimensional embeddings computed for each word.

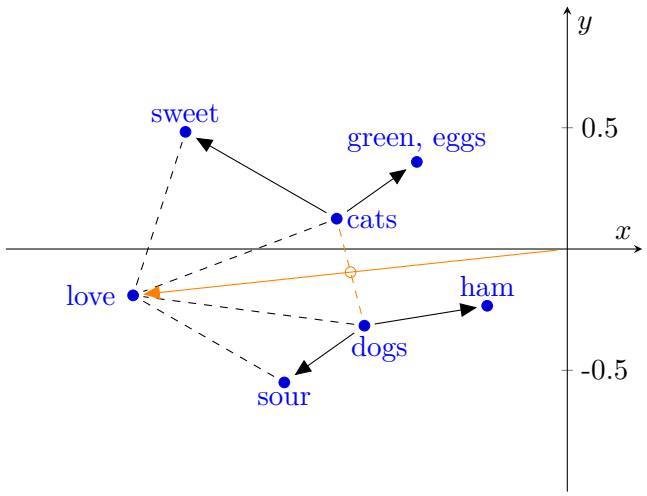



Figure 12: A plot of the two-dimensional vectors generated for the 8 unique words in the Green Eggs and Ham corpus.

Since the spaces produced by these word embedding algorithms are vector spaces in a formal mathematical sense, we can use powerful tools from linear algebra to *re-orient* our visualization from Figure 12, allowing us to build our intuition about various aspectsof the corpus' semantic relations. For example, looking at the original plot, we can observe that the $x$-axis neatly separates *most* of the vectors into a cat quadrant (above the $x$-axis) and dog quadrant (below the $x$-axis), but that

the arrangement is a bit skewed such that the vector for "love" lies within the dog quadrant, despite the fact that it actually perfectly bisects the angle formed by the "cat" and "dog" vectors (that is, the orange line in the figure passes through the midpoint of the dashed orange line joining "cats" and "dogs", the point labeled in the figure via an unfilled orange circle). Thus, as researchers, we may be able to get an even better sense of the dog vs. cat relation by *rotating* the points clockwise until the vector for "love" is exactly aligned with the $x$-axis, as pictured in Figure 13.

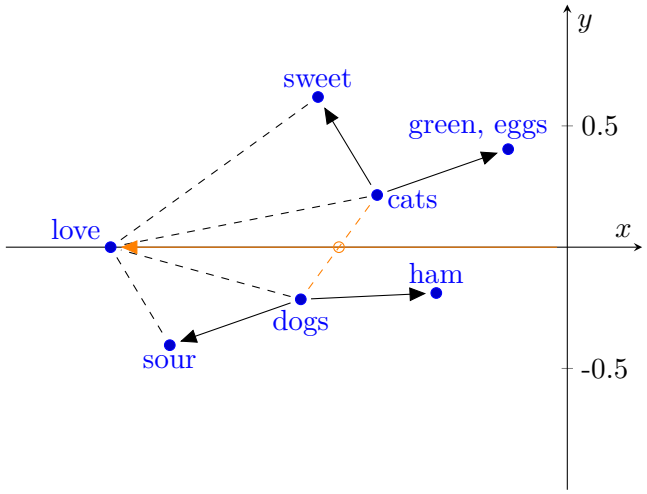



Figure 13: The same points as in Figure 12, but rotated clockwise so that the "love" vector is aligned with the $x$-axis.

By performing this rotation we thus obtain a clean separation which corresponds to our intuition with respect to the original corpus: words pertaining to cats in particular are in the cat quadrant (above the $x$-axis), those pertaining to dogs in particular are in the dog quadrant (below the $x$-axis), and the vector for the word "love", since it is neutral with respect to the cats-dogs split (i.e., since both cats and dogs are described as loving something), is now located exactly on the boundary point between the quadrants. Although the pattern which emerges could have been seen (via a tilt of the head) in the original plot, this ability to re-orient the embedding space visualizations as needed becomes crucial for more complex corpora.

As another example of a useful re-orientation, consider the case of a researcher who studies the rotated plot in Figure 13, identifies the dog quadrant vs. cat quadrant pattern, and wants to explore this axis further. By employing another linear-algebraic transformation we can rotate and

rescale the points so that the cats-to-dogs vector itself serves as our $x$-axis, allowing us to examine where the remaining points fall on this cats-to-dogs spectrum, as pictured in Figure 14. With this transformation, two salient splits now become evident: first, the cat and dog regions from before are now split (by construction) by the $y$-axis, with the cat region to its left and the dog region to its right. A second, new split can now be seen as well: that between food words (above the newly-constructed $x$-axis) and taste words (below the $x$-axis). The utility of this transformation is immediately apparent: with the four-quadrant separation resulting from the combination of these two splits (that is, starting from the upper-right quadrant and moving clockwise: dog-food words, dog-taste words, cat-taste words, and cat-food words), one could analyze a much more complex corpus by seeing how it "sorts" the words across the corpus into these categories. And indeed, computational linguist Dan Jurafsky performed just this sort of analysis on a much larger corpus—a collection of culinary descriptions from restaurant menus around the globe—as the basis for his award-winning book *The Language of Food: A Linguist Reads the Menu* (Jurafsky 2014).

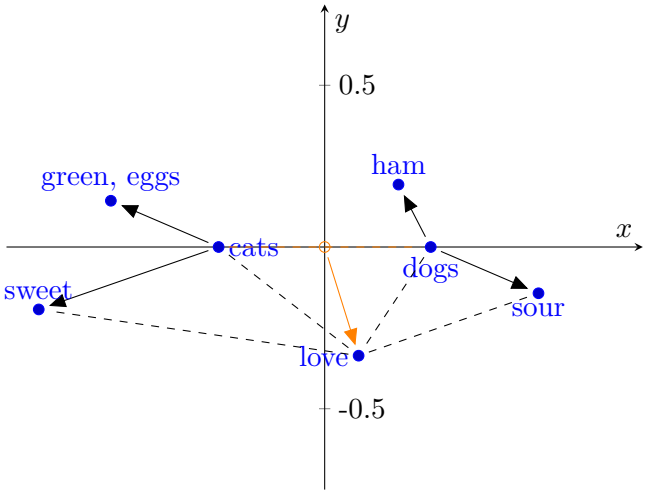



Figure 14: The same points as in Figures 12 and 13, but this time rotated and rescaled so that the vector from "cats" to "dogs" now serves as our $x$-axis.

A key lesson can be drawn from this this example for the sake of studying political theory[25], namely, that the effectiveness of these embedding algorithms for capturing semantic relations in

[25]Here we focus on political theory in general (i.e., viewed synchronically) rather than the history of political thought (viewed diachronically) since a genuinely historical study will require a model of how these embeddings evolve through *time*, which we indeed develop starting in Section 3.5 below.

a text depends upon how *systematically* the text employs key terms (i.e., whatever terms the researcher aims to study). That is, the algorithm's success in inferring meaningful semantic relations from just the four sentences in our constructed corpus was aided by the fact that the sentences straightforwardly delineated these semantic relations: each sentence contained a single subject-verb-object triad, with the subjects and objects shared across sentences, enabling the straightforward construction of the network structures visualized in Figures 12 through 14.

With this in mind, then, the most direct analogue of our hypothetical example within political theory would thus be a political-theoretic text which similarly aims to systematically discuss key terms and their interrelationships. The obvious candidate in Western political thought would be Hobbes' *Leviathan*, a text so self-consciously systematic that even a champion of contextualism like Skinner agrees with his Straussian critics in deeming it "probably the most plausible candidate" for "a classic text which 'makes eminent sense on its own'" (Tully 1988, p. 104, quoting Parekh and Berki 1973, p. 173)[26]. *Leviathan* stands out from other texts, that is, in Hobbes' attempt to construct an entirely self-contained and internally-cohesive system of political terms, making minimal reference (polemical or otherwise) to other texts[27].

Thus, in Figure 15, we present a two-dimensional visualization of an embedding space constructed from the original English text of Hobbes' *Leviathan*[28]. While perhaps not as cleanly-structured as the manually-constructed Green Eggs and Ham example, we *do* observe the algorithm distinguishing important clusters of words: key terms constituting the religious aspects of the text are seen in the bottom-right corner ("christ", "holy", "lord"), while a set of relevant vocabularic dyads can be seen throughout the space ("obey" and "fear", "civill" and "law", "common-wealth" and "soveraign", "authority" and "power"). The algorithm's formation of these clusters is already

[26]The strength of the claim can be seen even more clearly by considering the full passage from which the quoted snippet is taken: "unlike [Paine's *Agrarian Justice* or Burke's *Reflections*], *Leviathan* makes eminent sense on its own, [...] circumscribed but not limited by its own historical conditions." (Parekh and Berki 1973, p. 173)

[27]Hobbes' project in *Leviathan* thus stands in stark contrast to e.g. Marx's in *Critique of Hegel's 'Philosophy of Right'* (discussed in the next chapter), which consists entirely of explicit references to excerpts from Hegel, or his *Herr Vogt*, which is infused with dozens of implicit references in the form of puns and ironic statements which are unintelligible without knowledge of Vogt and his background.

[28]An interesting comparative study—outside the scope of our demonstrative example, but pertaining directly to some of Skinner's arguments in Skinner 2008—could be performed by training an embedding algorithm on both the original English and the later Latin edition of the text, to test whether any new vocabularic innovations were introduced by way of the latter.

encouraging, and perhaps surprising, given the fact that the standard applications of these word embedding algorithms in Computer Science typically train them on enormous datasets with billions of sentences[29]. To understand how such a generated space could be used as a tool for historical, sociological, economic, or other social-scientific research, we now turn to a review of existing works which have made effective use of embeddings to study social-scientific phenomena. We emphasize throughout, however, some key limitations of these works which our approach—by utilizing a newer generation of explicitly *context-sensitive* embedding algorithms—is able to overcome.

Since the introduction of the first widely-used word embedding algorithm, `word2vec` (Mikolov *et al.* 2013), in 2012, researchers in the social sciences have used these algorithms to incorporate information from textual corpora into studies which previously were restricted to using numeric or qualitative data. Recent studies have found, for example, that contextual embeddings are able to capture salient properties of social class (Kozlowski *et al.* 2019), the ideology of political manifestos (Rheault and Cochrane 2020), and the influence of economics on legal decisions (Ash *et al.* 2017). Of these three, the latter comes closest to our work, in attempting to study the linguistic properties captured by word embeddings using econometric methods for estimation of time-series effects.

A related literature, which predates the creation of word embedding algorithms, aims to quantitatively capture the existence, direction, and magnitude of ideological influence directly. Barron *et al.* 2018, for example, studies ideological influence across a time series of French Revolutionary debate transcripts by introducing "transience" and "novelty" metrics, which quantify how much the content of a given text is adopted by future texts, and how much it differs from the content of earlier texts, respectively. Unlike the previously-mentioned studies, however, this literature has yet to explore the new possibilities opened up by the development of semantic embedding algorithms, instead opting for the more well-established approach of probabilistic topic modeling (described here in Appendix A). Although these topic modeling algorithms are an effective tool for summarizing a corpus at a high level, they are ill-suited for the task of tracing out the trajectory of *particular*

[29]Google's LaBSE model for example, mentioned above, is trained on the CommonCrawl corpus, which contains over 6.4 PB (PB stands for "petabytes", with one petabyte being equal to 1,000 gigabytes (GB)) of data. The text file for Leviathan is about 1 kilobye (KB), meaning that the algorithm was able to detect these types of semantic relations even in a corpus 6,400,000,000,000 (6.4 trillion) times smaller than the standard CommonCrawl corpus used to train this algorithm in industry settings.

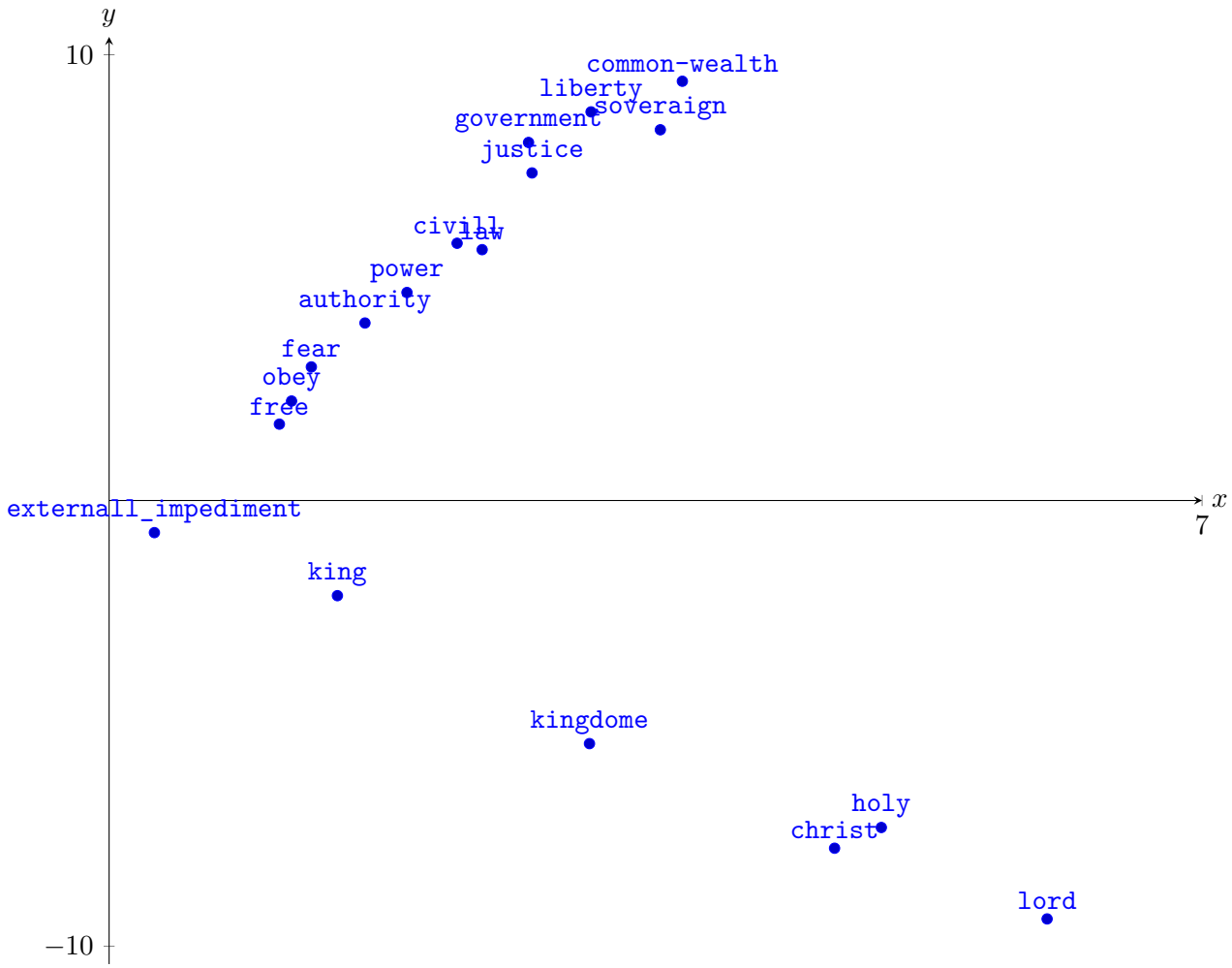



Figure 15: Visualization of a 2D embedding space generated from the English text of Hobbes' *Leviathan*.

terms or concepts over time. Since we are interested in how Marx was able to cement his particular set of terms as *the* vocabulary for later socialist discourse, embeddings are uniquely effective in allowing us to look at exactly which contexts a given term was employed in by Marx, and how this differs from the term's typical contexts before and after Marx's intervention. In other words, while topic-modeling-based approaches can tell us *that* Marx's writings influenced future socialist discourse, embedding-based approaches can tell us *how* this influence operated – which terms or concepts Marx utilized in particularly novel ways, and which of these illocutionary moves were and were not effective in terms of influencing subsequent socialist thought.

Studies across both of these literatures, moreover, have yet to adopt the contextual *sentence* embeddings we use herein, which utilize a more recently-developed embedding method from 2019 known as `BERT` (Devlin *et al.* 2019). `BERT`, in essence, differs from previous embedding algorithms in that it generates *joint* numeric representations for words *and* the contexts in which they appear, rather than associating each word with a single embedding vector. This means that, e.g., the word "`bank`" within the sentence "`I took my money to the bank.`" would be given a different embedding from the same word within the sentence "`I took a nap on the bank of the river.`". With these context-differentiated vectors, then, we have a powerful computational tool for grappling with a key difficulty in linguistically-informed historical studies: the fact that "words denote and are known to denote different things at the same time." (Pocock 1985, p. 30)

As visualized in Figure 16, this approach to polysemy changes the unit of analysis in our model: we are no longer modeling words themselves, but rather *word-context pairs*. In this way, the context is *explicitly integrated* into the numerical representations we use to understand an author's use of language, providing researchers with a quantitative implementation of precisely the type of Cambridge School-style contextual approach described above.

This improvement upon the original set of word embedding algorithms is therefore crucial, we contend, for capturing the nuanced uses of language which occur frequently in political-ideological polemics—the same concern which originally motivated the Cambridge School shift from what an author is saying to what they are *doing* with their words. Marx's 1860 polemic against Karl Vogt (Marx 1860), for example, makes use of several puns and purposeful misspellings of the names of those he is attacking, thus evoking contexts which the reader would not otherwise have read into the "standard" usage of the words. For example, he calls a particular political opponent with the surname Ranigel "Ran-Igel", likening him to an "Igel", the German word for hedgehog—we would therefore want to keep Marx's use of "Igel" in this sense separate from uses of "Igel" in general German texts. In general, due to the harsh censorship of political writings under the regimes of Friedrich Wilhelm III and IV (1797–1840 and 1840–1861, respectively)[30], pre-`BERT` text-analysis methods which are unable to capture the wide variety of ironic, metaphorical, and figurative uses

[30]See Rose 1978 for a detailed treatment of the effects of this censorship on Marx's writings, and Prawer 1976 for a chronological examination of the explicit and implicit literary references in Marx's writings.

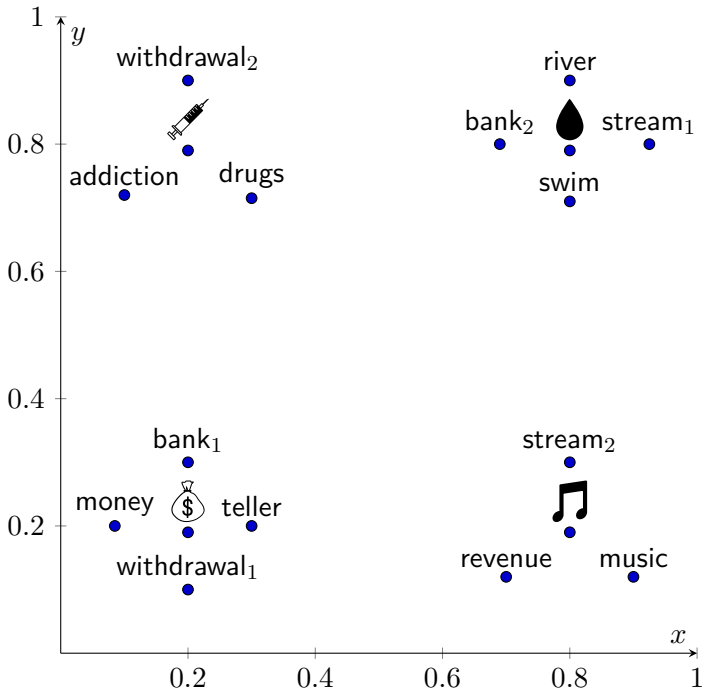



Figure 16: A visualization of BERT's ability to compute numeric representations of the same word in multiple different contexts.

of words would thus be unable to capture many important political illocutions throughout 19th century German political discourse, which had to be performed almost exclusively by way of these figures of speech.

### 3.2 Visualizing Contextual Fields

While the embedding spaces we've seen thus far have, for expository purposes, been *constructed* as two- or three-dimensional spaces, these toy models are insufficient in practice for analyzing the semantic content of a corpus. The rich variety of meanings which a set of words can be infused with—even within a single text, much less across a collection of texts in dialogue with one another—cannot be arranged neatly along two or three axes. For this reason, even the most bare-bones embedding algorithms typically aim to embed words and phrases within an extremely high-dimensional space: 300 dimensions in the original word2vec algorithm, for example, or 768 in the case of BERT.

The tradeoffs between lower- and higher-dimensional spaces can be seen by comparing the left

and right sides of the decomposition equation shown in Figure 9. Each value in the right-side matrix must be estimated from the corpus information encoded in the left-side matrix. In the terminology of statistics, each additional term that must be estimated "costs" an additional *degree of freedom*, which can be thought of as a measure of the "richness" of the information provided by the corpus. If the word "alienation" only appears in the corpus once or twice, for example, it will be difficult to generate a meaningful high-dimensional vector for this term, as this vector's semantic content is dependent upon (in fact, under the Firthian hypothesis, is *defined by*) the *range* of contexts in which the word is seen. In Figure 17, for example, we can see that word 4 appears only one time in the corpus, in a single context (context 5) in which no other words appear. In this case, therefore, `BERT` has no information about this word's relationship to any other word in the corpus, and thus no way to infer where its 768-dimensional vector (the fourth row in the Word Vectors matrix on the right-hand side of the equation, containing values $u_{4,1}$ through $u_{4,768}$) should be placed relative to the other word vectors in the generated space. In cases like these, we say that the dataset does not have a sufficient level of *statistical power* for estimation of embeddings via this method[31].

For intuition regarding this tradeoff in language modeling, consider Quine's *gavagai* thought experiment described above. If we have only one or two observations of a foreign-language speaker saying *gavagai* while pointing at a running rabbit, we won't have a rich enough set of contexts for this term to be able to differentiate whether, e.g., it refers to the rabbit itself, to the rabbit's fur, to a running animal in general, to creatures with tails, the act of running itself, etc. With the few linguistic "degrees of freedom" available to us at this point, we can only at best infer some vague properties of the term—perhaps that it has to do with motion more so than stillness[32]. As we observe the term being used more and more across a wider and wider range of contexts, however, we start accumulating enough "linguistic degrees of freedom" to allow us to more and more confidently and precisely map *gavagai* within the broader linguistic field[33] of the foreign language.

[31] We note that this term is used *in relation to* the method since, for example, the Word-Context matrix in Figure 17 *could* have a sufficient level of statistical power for estimation of much smaller (say, two- or three-dimensional) word vectors.

[32] Stretching the metaphor a bit for clarity: if our goal was to generate a mental word embedding space for the new language, after these few observations we could (at most) conservatively move the *gavagai* point slightly closer to the point for *motion* (or more accurately, given the inner workings of word embedding algorithms, the point(s) for terms which our model hypothesizes as highly likely to represent or relate to the concept of motion).

[33] In Section 3.5 below we will distinguish more precisely between linguistic *utterances*—the particular instances

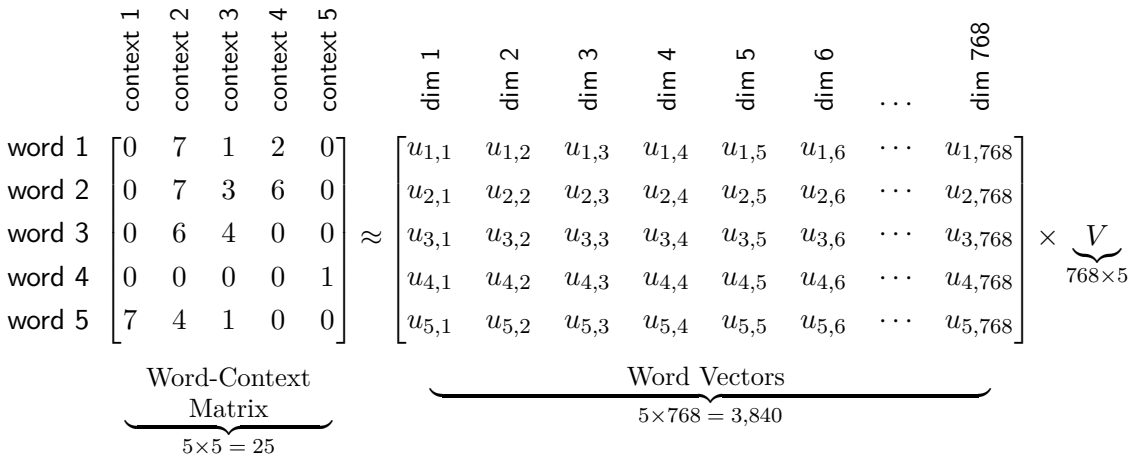



Figure 17: A case where the information about the corpus (the Word-Context Matrix on the left side) does not have sufficient statistical power to allow estimation of the word and context vectors (the matrices on the right side). At a high level, the algorithm is unlikely to be able to draw meaningful estimates for 7,680 parameters—the 3,840 entries in the word vectors matrix (the $u$ parameters), plus another 3,840 entries (not labeled) in the context vectors matrix $V$—using only 25 input data points. Looking more closely at the *structure* of the input data, we also see that it has no information from which to infer word 4's context-sensitive relation to other words, as it appears in only one context which is not shared by any other word in the corpus.

Once we *have* accumulated enough observations from the corpus to feel confident in our estimated embedding vectors, however, a second key question arises: how can we *use* this high-dimensional embedding space to draw meaningful inferences about the language of different authors or across different timespans within our corpus? Turning to the high-dimensional vectors produced by the embedding algorithm, we may ask, what good are (e.g.) points in a 768-dimensional geometric space when humans can only visualize points in at most two or three dimensions? While we do have tools like the similarity metrics discussed in Section **??** above, in a sense these only provide hints as to the full landscape of the embedding space[34]. We can do better using a set of visualization algorithms created specifically to construct 2D and 3D representations of high-dimensional spaces.

where we've observed *gavagai* being used—and the broader linguistic *field* from which these utterances were constructed (the Saussurean *parole* and *langue*, respectively). For now it suffices to understand "linguistic field" as a working model of the foreign language which is built using data in the form of utterances of *gavagai*.

[34]To see this, consider e.g. the task of trying to reconstruct a subway map given only the distances between stations. As illustrated for the Swiss rail system in Dokmanic *et al.* 2015, even sophisticated algorithms can only produce an approximation of the underlying space (this process is nonetheless crucial for e.g. inferring 3D protein structures; see Vendruscolo *et al.* 1997)

The $t$-distributed Stochastic Neighbor Embedding ($t$-SNE) algorithm, for example, constructs a 2D (or 3D) plot with the goal of capturing the "neighborhood structures" present within a higher-dimensional space, identifying distinct clusters of points which are closer to one another than to the other points in the space. For intuition, one can imagine taking a point in 3D space $\alpha^{\ominus}$ and asking, if I construct a small sphere centered at this point and grow it larger and larger, how soon will it come into contact with another point in the 3D space $\beta^{\ominus}$? The $t$-SNE algorithm constructs corresponding points $\alpha^{\circ}$ and $\beta^{\circ}$ in *2D* space such that, if I construct a small *circle* centered at $\alpha^{\circ}$ and grow it larger and larger, a similar property will hold with respect to how long it takes for that circle to come into contact with $\beta^{\circ}$. In this way, nearby points in $N$-dimensional space remain nearby in the two-dimensional visualization.

A separate visualization tool called Principal Component Analysis (PCA) also constructs low-dimensional representations of high-dimensional spaces but, unlike $t$-SNE, focuses on capturing the *variance* among points rather than their neighborhood structure. To this end PCA can be thought of simply as projecting all the points in the space down to two dimensions, from every possible angle, then choosing the most interesting of these projections with respect to how *spread out* the points are. Thus, taking the scatterplot of points in Figure 18 as an example, we can see in the figure what these points in 2D look like when projected down into one dimension, whether onto the $x$-axis (the short lines along the horizontal axis at the bottom) or the $y$-axis (the short lines along the vertical axis on the left).

However, this figure also illustrates a potential drawback of these "standard" projections onto the $x$-axis or $y$-axis, namely, that they may not in fact be the most interesting projections with respect to capturing the directions along which the points vary most. If the points represented word embeddings learned from the WPA Slave Narrative Collection (Yetman 1967), for example, the reason for these points' diagonal arrangement is likely to be far more interesting to researchers than the points' original $(x, y)$ coordinates. Perhaps points closer to the bottom-left of the plot are those used in the context of describing masters, and those closer to the top-right are those used when describing fellow slaves, for example, while points closer to the top-left corner are more likely to refer to women than those closer to the bottom-right corner, regardless of their placement on

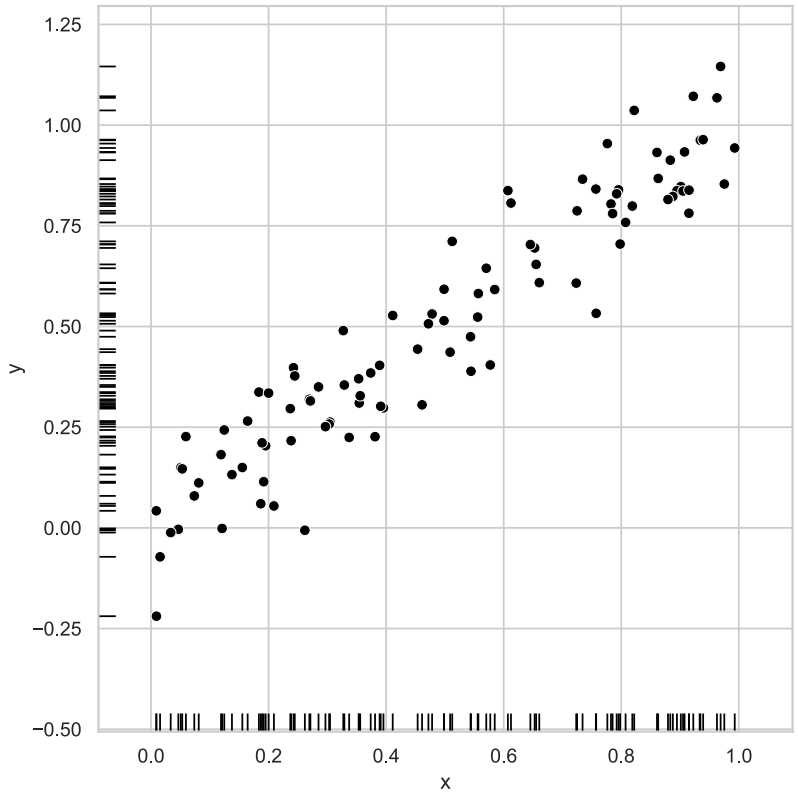



Figure 18: 100 randomly-generated points in 2D space, showing potential projections of these points down into one dimension on both the $x$ and $y$ axes (the short vertical and horizontal lines, respectively).

the main diagonal.

The PCA algorithm captures this intuition, that the directions of highest *variance* may be most salient to the researcher, by tossing out the original $(x, y)$ coordinate system and replacing it with a coordinate system based on Principal Components: the First Principal Component axis is the axis along which the data varies the most, while the Second Principal Component axis is the direction along which the data varies second-most. This process can be continued, in fact (e.g., to generate a 3D PCA plot), up to the $N$ dimensions of the original set of points, revealing its power as *both* a dimensionality-reduction *and* data-reorientation tool. In our working WPA Slave Narrative Collection example, for instance, our PCA plot has the same dimensionality (2D) as the original plot, but re-orients the points so that the $x$-axis is the axis of maximum variance and the $y$-axis that of second-highest variance, as illustrated in Figure 19.

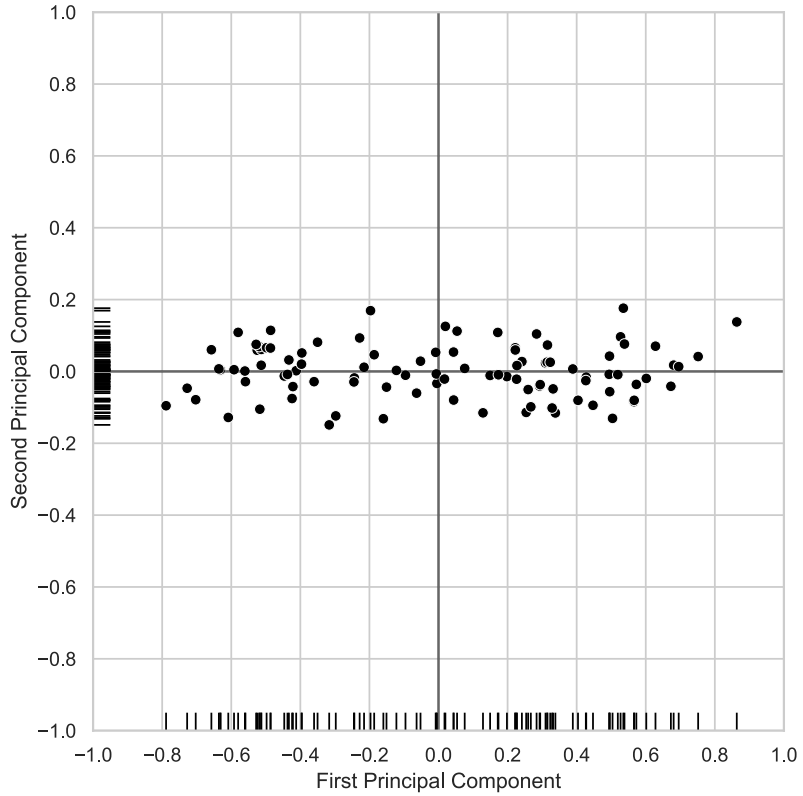



Figure 19: The same data as plotted in Figure 18, but transformed via PCA so that the horizontal axis represents the direction of greatest variance (the First Principal Component) while the vertical axis represents the direction of second-greatest variance (the Second Principal Component).

To see the usefulness of this method in practice, we start by computing two-dimensional word embeddings for terms in the WPA Slave Narratives corpus, obtaining the space illustrated in Figure 20. And indeed we see the type of non-$(x, y)$-oriented pattern that we hypothesized earlier: isolating key terms for race ("white" and "black") and status within the slave system ("slave" and "master"), we can observe that a move from bottom-right to top-left corresponds to a move from "black" to "slave", and that a nearly identical move in the same direction brings us from "white" to "master". With this observation in mind, then, we can choose our own horizontal and vertical axes and then analyze other terms *with respect to* this newly-meaningful left-right or up-down orientation. For example, we generate Figure 21 with the same data but with the horizontal axis now representing the white-to-black orientation we observed in the original plot.

Interestingly, after performing this translation, we can observe a second emergent feature of the word vectors' orientation: although in the original plot the move from "woman" to "man" seemed to be broadly similar to the move from "white" to "master" or "black" to "slave", once we differentiate the latter pairs by choosing them to be the horizontal and vertical axes (respectively), we can see that in fact the "woman" to "man" direction is *diagonal* with respect to this new orientation—as we observed in our theoretical example above, this indicates that perhaps our analysis could benefit even more from setting the vertical axis to represent gender, as illustrated in Figure **??**.

The arbitrariness of the $(x, y)$ coordinates, and thus the impetus for employing PCA as one possibility for deriving *meaningful* coordinates instead, is important enough to re-emphasize here: since the word embedding algorithms don't know the particular types of meaning we hope to capture, the coordinates generated for each word have no *a priori* human-interpretable meaning at all; they are chosen solely on the basis of placing words closer together or further apart based on the similarity or difference between the contexts they appear in. In a 2D space, for example, if *gavagai*'s initial embedding (typically randomly-chosen) lies directly below that of *bird* and directly above that of *fish*, but the algorithm observes *gavagai* being used in a context which shows it is neither of these animals, the only way to increase its distance from *both* points is to move *gavagai*'s embedding left or right, i.e., to modify its $y$ coordinate. Since the same logic holds in cases where the three points randomly start out e.g. in a horizontal line (*gavagai*'s embedding would have to

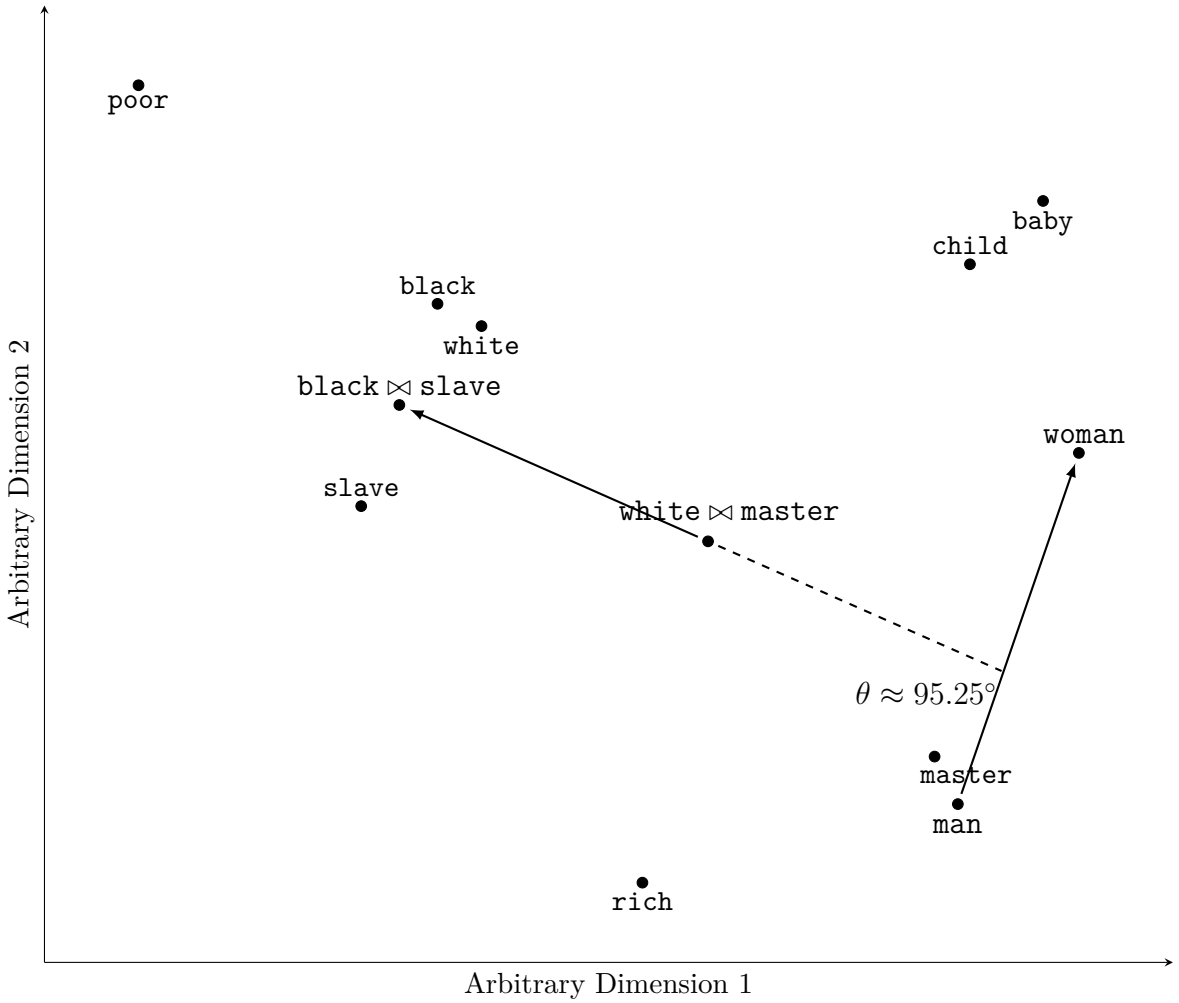



Figure 20: A two-dimensional embedding space trained on the WPA Slave Narratives corpus, where we draw solid arrows connecting points for key racial terms ("white" and "black") and for terms denoting status within the slave system ("slave" and "master"), along with a dashed arrow connecting points for gendered terms ("man" and "woman").

be moved up or down), or on the $x$-$y$ diagonal (its embedding would have to be moved along the off-diagonal), we can see that the coordinates themselves have no stable or human-interpretable meaning on their own. It is up to the researcher to choose—on the basis of their research goals and/or intuitions regarding their field of study—how to transform the data to make sense of the embedding algorithm's outputs. Thus we have another example of how, as emphasized above, these algorithms do not *automate* social-scientific research, but *do* provide a powerful addition to the researcher's toolbox, to be used in conjunction with their own domain knowledge and expertise.

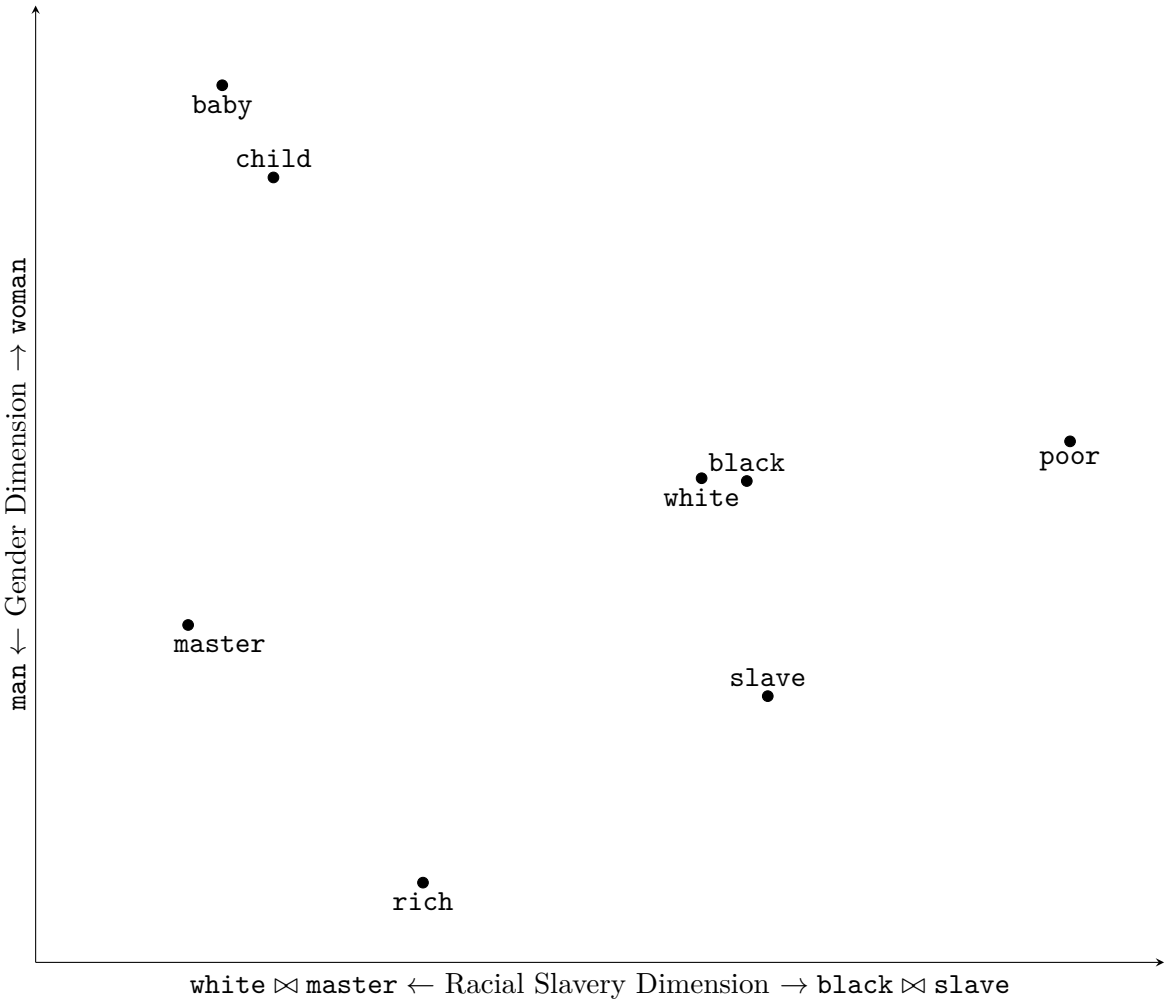



Figure 21: The same points displayed in Figure 20, but transformed so that the horizontal axis represents the white-to-black direction within the space (`white` ⋈ `master` on the left, `black` ⋈ `slave` on the right), and the vertical represents the man-to-woman direction (`man` below, `woman` above).

There is an unexplored dimension of the WPA embedding examples thus far, however, which may be crucial for a researcher's understanding of the corpus, especially in the social sciences and humanities. Namely: the spaces we've generated to this point have produced a *single* point for each word, by *averaging* its uses in various contexts over *all* the interview subjects, when in reality what social science researchers often care about is precisely how the uses of these words *differ* between different individuals or groups. If a researcher is analyzing the WPA Narratives with the goal of understanding 19th and early 20th century notions of freedom in the U.S., for example, it would be

crucial to differentiate between how the term was used by slaves themselves versus how it was used by masters or overseers, just as historians have analyzed how (e.g.) John C. Calhoun and Frederick Douglass infused terms like "liberty" and "freedom" with very different semantic contents.

To this end, in the next section we describe a set of algorithms developed specifically to model not only the different *senses* in which a word is used (which we saw above in our explanation of the `BERT` algorithm), but also how different *authors* emphasize or de-emphasize these senses. This brings us one step closer to our goal of understanding the history of political thought as a history of illocutionary moves since, as we argue, these author-specific emphases and de-emphases can be seen as textual manifestations of their attempts to *steer* discourse in a certain direction: to steer popular understandings of "liberty" towards what we would today call "negative liberty" (freedom as non-interference) in Hobbes' case, for example, or to steer socialist discourse away from utopian ideals and towards political-economic categories, in Marx's.

### 3.3 Author-Specific Embedding Spaces

As we gestured towards at the end of the previous section, in order to further pinpoint the pathways through which political thought has evolved we'll need a model which can keep track not only of the different *senses* in which key terms are used, but also of the particular *individuals* who use these terms in certain ways which shift their usage going forward. A recently-developed extension of the original `BERT` algorithm (Welch *et al.* 2020) provides us with exactly the tool we need in this case, in explicitly constructing an additional personalized embedding for each (word, author) *pairing* on top of the main contextual embedding for each word irrespective of author (i.e., relative to the entire corpus). This enables us to capture not only the shared terms and figures of speech available to all authors within a discursive community, but also the *particular* ways in which they are employed—or wielded—by individual authors. Indeed, if one accepts (as we do) Skinner's contention that historians ought to treat these terms and figures of speech "less as statements about the world than as tools and weapons of ideological debate" (Skinner 2012, p. 177), these author-specific embeddings provide a crucial tool for explicitly modeling how particular word-weapons were wielded by particular actors in particular wars of ideas, such as the battles for hegemony over

socialist discourse that we analyze in future chapters.

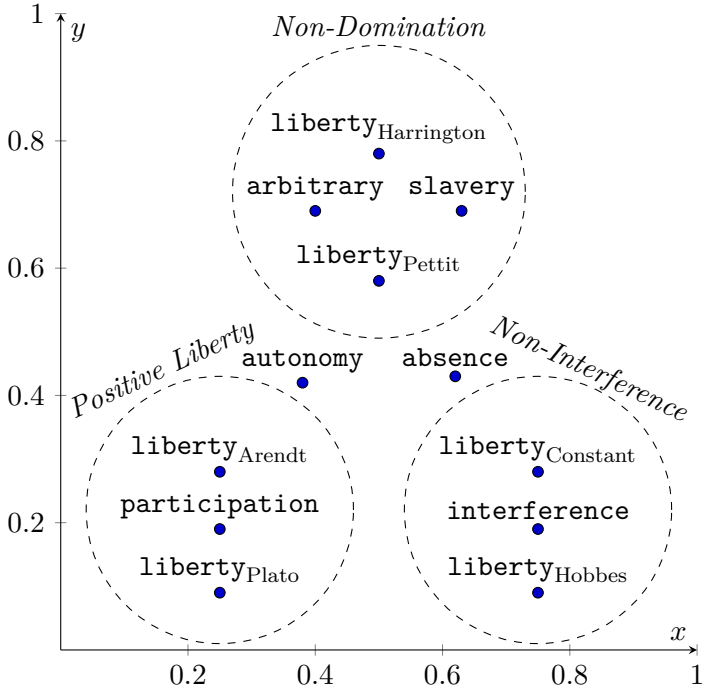



Figure 22: A visualization of BERT's ability to model the different contexts (and thus, senses) in which words are used by different authors. In this case interference falls squarely within the negative liberty (liberty as non-interference) cluster, as a term central to distinguishing negative from other forms of liberty, whereas absence falls between two clusters as it can be employed in both republican ("absence of arbitrary power") and negative-liberty ("absence of interference") contexts.

Importantly, this method does not generate a separate embedding *space* for every author, since our entire goal is to be able to *compare* different authors' vectors on some common scale[35]. Instead, in the author-specific BERT estimation, each author's specific vector $\overrightarrow{w_i}$ for a given term $w$ is estimated as an *offset* relative to the vector for $w$ in the MAIN vector space, $\overrightarrow{w_{\texttt{MAIN}}}$. Given a vector $\overrightarrow{w_{PE}}$ representing the centroid of political-economic discourse within the broader ideological vector space (estimated via a procedure we detail in the next section), for example, this allows us to instantly check whether an author $A$ tends to use a term $w$ in a more political-economic context than some other author $B$, by checking whether $\mathsf{dist}(\overrightarrow{w_A}, \overrightarrow{w_{PE}}) < \mathsf{dist}(\overrightarrow{w_B}, \overrightarrow{w_{PE}})$, or relative to the "average"

[35]i.e., for authors $A$, $B$, and $C$, the distance $\mathsf{dist}(\overrightarrow{w_A}, \overrightarrow{w_B})$ between $A$'s vector for some word $w$ and $B$'s vector for $w$ must be on the same scale as the distance $\mathsf{dist}(\overrightarrow{w_A}, \overrightarrow{w_C})$ between $A$'s vector for $w$ and $C$'s vector for $w$, as well as the distance $\mathsf{dist}(\overrightarrow{w_B}, \overrightarrow{w_C})$ between $B$'s vector for $w$ and $C$'s vector for $w$.

usage of the term across the entire corpus, by checking whether $\mathsf{dist}(\overrightarrow{w_A}, \overrightarrow{w_{PE}}) < \mathsf{dist}(\overrightarrow{w_{\texttt{MAIN}}}, \overrightarrow{w_{PE}})$.

Thus, while studies like Kozlowski *et al.* 2019 are able to discover shifts in the overall discourse around social class during the 20th century, for example that different education-related terms tended to become more dichotomous along the upper-class/lower-class axis, this modification of `BERT` would allow them to identify *which authors in particular* were ahead of the curve, adopting the newer senses of education-related terms earlier than others. Although a fully-developed methodology for testing the *causal* impact of these early-adopters on the broader discourse is outside the scope of this paper, we are still therefore able to identify a necessary condition for causal influence—temporal precedence—and thus develop plausible causal hypotheses based on which authors' embeddings precede the overall embeddings in terms of the movement of key terms along socially meaningful axes.

To this end, especially when we turn to *diachronic* embeddings below, we keep in mind the notion of *Granger Causality* between two time-evolving data series $X_t$ and $Y_t$: if the values of $X$ up to some time $t$ help predict future values of $Y$ $(Y_{t+1}, Y_{t+2}, \ldots)$ better than one could predict them solely on the basis of past values of $Y$, $X$ is said to exert Granger causality on $Y$. Thus, for example, if we consider $X_{\leq 1848}$ to be data representing Marx's pre-1848 speech acts, $Y_{\leq 1848}$ to be data representing language use by socialists besides Marx before 1848, and $Y_{>1848}$ to be data representing language use by socialists besides Marx *after* 1848, we can say that Marx's speech acts exerted Granger causality on socialist discourse to the extent that knowing $X_{\leq 1848}$ helps us predict the post-1848 trajectory of socialist discourse better than we could using $Y_{\leq 1848}$ alone (i.e., pre-1848 socialist discourse excluding Marx). Abstracting away from the details, then, when historians of political thought make claims like "one cannot understand the trajectory of post-1848 socialist thought without understanding Marx [or Hegel, or Babeuf, etc.]"[36] they are (perhaps unknowingly) making claims of Granger causality[37].

While testing for this type of causality using these author-specific embeddings is therefore

[36]Sholomo Avineri's *The Social and Political Thought of Karl Marx*, for example, introduces its discussion of August von Cieszkowski with the claim that "one cannot fully grasp Marx without recourse to Cieszkowski." (Avineri 1968, p. 125)

[37]On the applicability of this concept in the social sciences see the short entry in Lewis-Beck *et al.* 2003, p. 439, or the more in-depth discussion in Hlaváčková-Schindler 2012. Within the humanities as such we have been able to locate only one work, Wevers *et al.* 2020, which explicitly focuses on this conception of causality.

possible in theory, in practice the problem of moving from associational to causal analysis is typically overshadowed by an even more fundamental problem, related to our discussion of *degrees of freedom* in Section 3.2 above. While in that section we discussed how the researcher must have enough text—with words ranging over a rich enough variety of contexts—to allow estimation of all $N$ entries within each word's $N$-dimensional vector, the problem becomes even more acute when estimating author-specific spaces. Now, if there are $M$ authors in the corpus, the researcher must have enough text from *each author* to estimate all $N \times M$ entries across the $M$ vectors generated for each word, plus the additional $N$ entries in the $N$-dimensional "main" word vector. Methodologically, this is why the algorithm's developers chose to estimate the single $w_{\texttt{MAIN}}$ vector for each word, and then learn each author's specific word embeddings $w_a$ as an *offset* of this `MAIN` vector: this allows the algorithm to first pool *all* of the scarce information about $w$ across the entire corpus towards estimating $w_{\texttt{MAIN}}$, then to estimate the author-specific embedding $w_a$ on the basis of how many degrees of freedom remained (with respect to the word $w$) to "spend" on this additional estimation.

In practice this means that researchers must take care when interpreting cases where an author's embedding $w_a$ for some key word $w$ seems not to differ much from $w_{\texttt{MAIN}}$. While ideally—that is, given enough textual data from each author—this indicates that the author's usage of $w$ indeed did not differ substantially from the general usage of $w$ across the corpus, it could also mean simply that the algorithm did not have enough data on this author to evaluate their usage, and thus "guessed" their usage to be at or near the corpus-average usage, for lack of any better way to choose. This latter case corresponds to the heuristic where, for example, if we only know that a man is Dutch, our best guess as to his height would be just the average height of Dutch men (182.5 cm, or about 6 ft.).

This analogy also points the way towards a potential modification of the approach, however, which we use to minimize the chances that this collapse-to-the-average affects our results. Within the analogy, note that if we could obtain additional salient information about this man, such as his membership within a *subset* of the Dutch population, we could narrow the group over which we're taking the heuristic average. For example, if we learned that he was on the Dutch Olympic basketball team, we could do much better by guessing his height to be the average height of this

team's members (about 6 ft. 7 in.), rather than that of the Dutch population as a whole. Thus in our case, to address this collapsing-to-the-average issue for authors with a small number of extant texts, we instead group individual authors into "meta-authors" based on the discursive community they are generally associated with in the historical literature—chosen to be as specific and/or exclusive as possible to minimize the distortions introduced by algorithmic averaging[38]. Thus, for example, the texts of Bruno and Edgar Bauer, Arnold Ruge, Max Stirner, etc., are combined into one Young Hegelian meta-author in order to estimate a vector $\overrightarrow{w_{YH}}$ representing the centroid of Young Hegelian discourse (as defined by this author-to-group mapping) within the broader ideological space of 19th-century German discourse. Importantly, however, this approach is *not* used to generate the political-economic and Hegelian vectors which serve as our orthogonal basis vectors, for reasons we describe in the next section.

### 3.4 Discursive Fields as Embedding Clusters

In the previous section we introduced two special vectors $\overrightarrow{w_{PE}}$ and $\overrightarrow{w_H}$, representing the centroids of political-economic and Hegelian discourse respectively, which we do *not* compute via author-specific embedding estimation. Instead, to minimize the dependence (in the statistical sense) between our two basis vectors and the vectors like $\overrightarrow{w_{Marx}}$ for which we want to observe movement over time, we compute these basis vectors as centroids of word clusters which are derived independently via the $\mathsf{cTFIDF}$ measure[39], which generates a ranking of all terms in the corpus on the basis of how "unique" they are to political-economic texts relative to Hegelian texts (and vice-versa, by taking the $N$ terms with lowest, rather than greatest, $\mathsf{cTFIDF}$ scores).

[38] Considering the Dutch basketball player example again, this matches our intuition that the average height of the Olympic team would provide a better guess than, for example, the average height of all Dutch basketball players or athletes in general, as the former represents a much smaller height-related subset of the population.

[39] $\mathsf{cTFIDF}$ stands for the Class-Based Text-Frequency minus Inverse Document Frequency measure, formally defined via the formula

$$\mathsf{cTFIDF}(w_i, c_j) = \mathsf{cf}(w_i, c_j) \cdot \log\left(1 + \frac{\overline{N}}{\mathsf{f}(w_i)}\right),$$

where $w_i$ represents the $i$th word in the corpus, $c_j$ represents the $j$th class (grouping of documents), and $\overline{N}$ represents the average number of words per class. $\mathsf{cf}(w_i, c_j)$ measures the number of times word $w_i$ appears across class $c_j$'s documents, while $\mathsf{f}(w_i)$ measures the number of times $w_i$ appears across the entire corpus. In this section we provide intuition for how this formula is able to capture the uniqueness of a word $w_i$ to a class of documents $c_j$, relative to all other classes (which we collectively denote as $c_{-j}$).

The intuition we aim to capture with this measure is that a word $w_i$ is more "unique" to a class of documents $c_j$ to the extent that it appears frequently *within* $c_j$'s documents, but *infrequently* within the documents of other classes. In the corpus of nineteenth-century socialist texts we use in later chapters, for example, we may observe the German word `das` appearing frequently in Marx's writings, but it is not *unique* to Marx's writings, since it also appears frequently in the writings of the other authors in the corpus. In this case, `das` has a high *class frequency* value relative to the class $c_{\mathsf{Marx}}$ representing the collection of Marx's writings (denoted $\mathsf{cf}(\texttt{das}, c_{\mathsf{Marx}})$), but also a high *corpus-wide* frequency value (denoted $\mathsf{f}(\texttt{das})$). However, if we consider the term `Mehrwert` (surplus value) instead, the frequency with which this word appears in Marx's writings is much higher than the average frequency of this word across all nineteenth-century authors, so that $\mathsf{cf}(\texttt{Mehrwert}, c_{\mathsf{Marx}})$ is high but $\mathsf{f}(\texttt{Mehrwert})$ is low. Thus, as a simple measure of uniqueness, we can simply consider the ratio of these two quantities:

$$\mathsf{Uniqueness}(w_i, c_j) = \frac{\mathsf{cf}(w_i, c_j)}{\mathsf{f}(w_i)},$$

so that

$$\begin{aligned}\mathsf{Uniqueness}(\texttt{das}, c_{\mathsf{Marx}}) &= \frac{high}{high} \approx 1, \text{ while} \\ \mathsf{Uniqueness}(\texttt{Mehrwert}, c_{\mathsf{Marx}}) &= \frac{high}{low} > 1,\end{aligned}$$

matching our intuition regarding the uniqueness of a word to a particular author. Note that the formula also allows us to measure how *rare* a word is for a particular author: if Marx uses the term `Utopie` significantly *less* often than the other authors in the corpus, we'll now have a value for $\mathsf{cf}(\texttt{Utopie}, c_{\mathsf{Marx}})$ which is lower than $\mathsf{f}(\texttt{Utopie})$, giving us

$$\mathsf{Uniqueness}(\texttt{Utopie}, c_{\mathsf{Marx}}) = \frac{low}{high} < 1,$$

so that now we can qualitatively characterize an author's use of a word $w_i$ in three ways:

- If $\mathsf{Uniqueness}(w_i, c_a) < 1$, the word is used *less* often by the author than its average usage

across authors in the corpus

- If $\mathsf{Uniqueness}(w_i, c_a) \approx 1$, the author uses the word $w_i$ at about the average rate it is used by all authors in the corpus, and
- If $\mathsf{Uniqueness}(w_i, c_a) > 1$, the author uses the word $w_i$ *more* often than its average usage across all authors in the corpus.

The formal cTFIDF formula, then, is simply this intuitive Uniqueness measure with one modification enabling it to better capture a key empirical property of "natural" (human) languages: namely, that word frequencies in these languages exhibit what's called a "power law" distribution. In a linguistic context, this "law" is really just an observed tendency, found in nearly all known human languages, for the frequency of words to be inversely proportional to their frequency *rank* in the language. In other words, if the frequency of the most common word is $f_1$, the second most frequent word tends to occur 1/2 as often, the third most frequent word 1/3 as often, an so on, such that the frequency of the $n$th most frequent word can be written as $f_n = (1/n) \cdot f_1$.

In English for example, the most frequent word ("the") occurs about $f_1 = 50,000$ times per million words, the second-most frequent ("be") occurs about half as often ($f_2 = (1/2)f_1$), the third-most frequent ("and") about one-third as often ($f_3 = (1/3)f_0$), and so on, with this pattern (of the $n$th most frequent word occurring $1/n$ as often) continuing all the way down to the rarest words in the language. Given this tendency, then, the frequencies of words will exhibit massive (exponential) differences when comparing among high-frequency words, but tiny differences when comparing low-frequency words, as shown by the orange points in Figure 23. In the case of our Uniqueness measure, however, we don't want the difference between (e.g.) the first and second most frequent words to count 100 times more than the difference between the 100th and 101st most frequent words. Thus, to prevent this linguistic property from skewing our measurements, we replace the denominator with its *logarithm*, ensuring a more gradual decrease in the denominator's value as we measure uniqueness for less frequent terms, illustrated by the blue points in Figure 23.

With this modified, linguistic-theory-informed version of our Uniqueness measure in place we can to capture, in a mathematically- and linguistically-principled manner, the *centrality* of certain

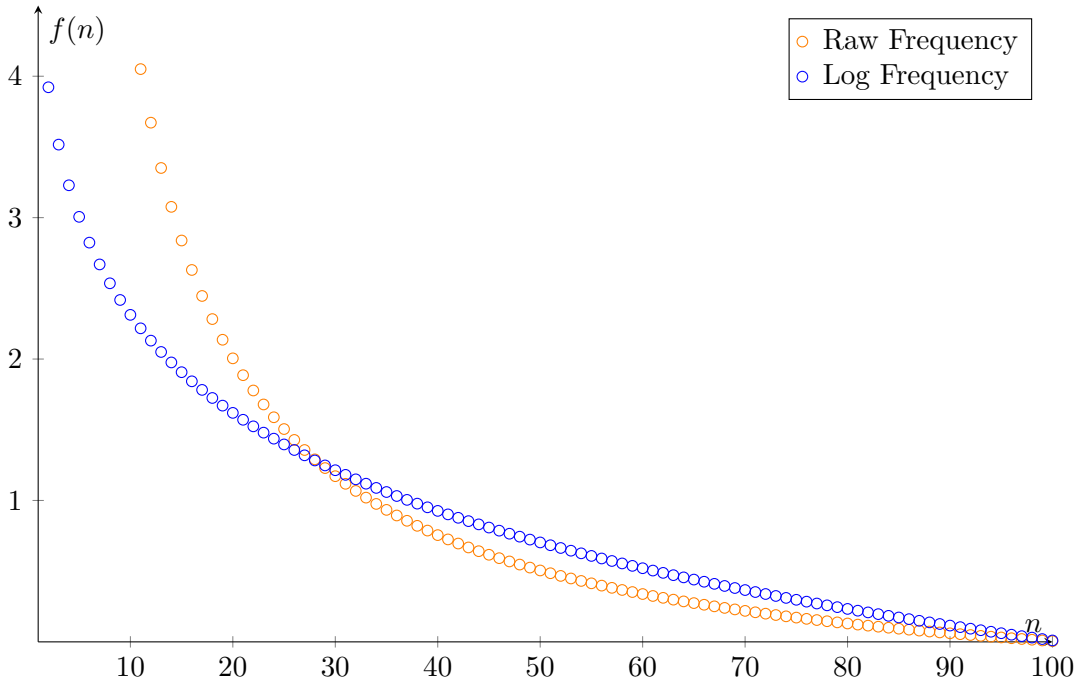



Figure 23: An illustration of how the logarithm (plotted in blue) "smoothes out" the original corpus-wide frequencies (plotted in orange), where the horizontal axis represents a word's ranking in the corpus ($n = 1$ representing the most frequent word, $n = 2$ the second most frequent, and so on) and the vertical axis represents the frequencies of these words, measured in both original and logarithmic units. In particular, note that the logarithmic plot decreases the magnitude of the differences between the first ~30 words, and then slightly magnifies the differences between the remaining ~70 words, relative to the original frequency plot.

terms to certain discursive genres: Hegelianism, political economy, Comtean positivism, etc. In future chapters we therefore use this approach to *anchor* our axes, giving us a stable coordinate system—a common ruler—with which to measure differences in word usage between authors as well as changes in usage over time within and across these broadly-defined genres.

Specifically, the construction of these axes proceeds as follows: given a corpus $C$ containing words $\{w_1, w_2, \ldots, w_n\}$ and a set of genres $\{G_1, G_2, \ldots, G_m\}$ (where each genre is associated with a subset of the texts in the corpus which the researcher believes to be sufficiently representative of that genre), the researcher can compute "genre-uniqueness scores" $u_{i,j} = \mathsf{cTFIDF}(w_i, G_j)$ for all (word, genre) pairings. With these scores in hand the researcher can then rank, for a chosen genre $G^*$, the $N$ words most unique to $G^*$ by choosing the $N$ words with the highest $\mathsf{cTFIDF}(w_i, G^*)$ values. Once word embeddings for all words in the corpus have been generated, then, a *centroid*

vector representing the genre $G^*$ within this space can be computed as the *weighted average* of the word vectors for the $N$ words found to be most unique to $G^*$. Once these centroid vectors have been computed, we can use them to orient ourselves within the otherwise arbitrary[40] geometric space. The movement of points through this space, for example, which before corresponded only to non-human-interpretable changes in the numeric values of a vector's 768 coordinates, can now be meaningfully interpreted with respect to one or more of these computed centroid vectors: a movement towards a centroid $G$ represents an increase in similarity to the semantics of this discursive genre, and vice-versa for a movement away from $G$.

Social-scientifically, this interpretive lens enables us to analyze the linguistic field in terms of *comparative statics*: holding some variables constant to enable the measurement of variation in others with respect to this static value. The "Hegelian-ness" of various authors in the corpus, for example, can be measured by comparing how their author-specific embedding vectors vary in relation to the static Hegelian genre-centroid vector. Or, we can evaluate an author's "Hegelian-ness" over time, by tracking changes in the distances between their author-specific vectors and the static Hegelian genre-centroid vector, over time.

This latter example, however, leaves some important questions unaddressed with respect to the nature of the linguistic fields we are studying: while treating a centroid for Hegel's writings as a constant in order to analyze changes in Marx may be reasonable, since these writings predated Marx's own (i.e., Hegel was not in interlocutory or back-and-forth dialogue with Marx in the same sense that Proudhon was), what do we do about the case where one author's speech acts feed into and have an effect on the very linguistic field from which another author subsequently draws when constructing their own speech acts? To incorporate *time* as a variable in our analyses, we need to be able to model these complex linguistic feedback effects in a non-trivial way. That is, we need to explicitly augment our static embedding model with a model of the *temporal relationships* between individual speech acts and the linguistic fields from which they are drawn.

As the first step in this direction, then, we begin the next section with an introduction to Saussurean structuralist linguistics. In particular, we discuss the distinction between *langue* and

[40] Arbitrary, that is, with respect to the original $x$ and $y$ coordinates generated by the embedding algorithm.

*parole* introduced by this approach, corresponding to the distinction between what we've been calling *linguistic fields* and *speech acts* thus far, and explain how this perspective gives rise to a working model of the *mutual interactions* between these two co-constitutive phenomena which we can carry over into our computational approach.

## 3.5 Synchronic and Diachronic Analysis: Understanding the *Langue-Parole* Distinction

Notwithstanding the progress in the field of semantic embeddings described at the end of Section 2.3, resulting in algorithms more and more capable of capturing meaning holistically across larger and larger linguistic units, there remain additional aspects of language—aspects crucial to social-scientific analysis—which we have not yet discussed and which will require a complementary set of tools to capture. To understand what the models thus far have lacked (and thus what additional tools we'll need to employ), however, we'll need to understand the Saussurean distinction between *synchronic* and *diachronic* approaches to language analysis.

Ferdinand de Saussure, in his *Course on General Linguistics*, differentiates between two perspectives one can adopt when analyzing language: in the the first, which he calls the *synchronic* perspective, the linguistic entities—texts, utterances, etc.—are treated as if they exist in a single moment in time (thus roughly corresponding to the notion of *cross-sectional* analysis in the sciences), so that one can consider their interrelationships within a stable, static field of linguistic meaning. In other words, in the synchronic mode, one is concerned with understanding a particular collection of utterances, but *not* with the dynamics by which these utterances, once issued, subsequently changed the field of language from which future utterances were generated (nor are they concerned with the other temporal direction: how past utterances shaped the linguistic field which gave rise to the utterances of interest). In Pocock's phrasing, however, the linguistic field that Cambridge School historians study is *inherently*, definitionally diachronic, and must be studied in a diachronic mode:

> "the diversity of linguistic contexts that went to determine what could be said but were at the same time acted upon by what was said." (Pocock 1985, p. 2)

The models we've discussed thus far have all been essentially synchronic. The embedding algorithms, for example, are trained with the sole objective of capturing the relationships between *words* and the *contexts* they appear in across a given corpus, regardless of the temporal order in which the word-context pairs appeared. In reality, however, if we want our computational models to be able to aid in Cambridge School-style context-sensitive historical inquiry[41], we should take heed of the inherently diachronic emphasis of e.g. Pocock's definition cited above: "a history of actors *uttering and responding*".

This Saussurean jargon may, at first glance, seem like a distinction without a difference: to obtain a diachronic model, one can simply train a sequence of synchronic models, one for each time period. We argue, however, that this misses the point of language modeling entirely: word embedding models are useful social-scientific tools because they implement a powerful theory of linguistic discourse—namely, the Firthian Hypothesis that we shall "know a word by the company it keeps" (Firth 1957, p. 11)—which enables researchers to see (via the generated geometric spaces) the contours of how particular authors use particular words to do particular things. By the same logic, then, we argue that a diachronic language model is only useful or relevant for social-scientists to the extent that it can similarly implement an analytically-powerful theory of language change. Obtaining a sequence of trained word-embedding spaces is therefore necessary but *not* sufficient for having a useful social-scientific tool. It must be paired with a theory regarding the *relationship* between these time-slices, just as pairing the outputs of word embedding algorithms with the Firth Hypothesis provides us with a social-scientifically-grounded interpretive tool for understanding the relationship between words in the embedding spaces.

To this end, in the next section we survey the political-theoretic literature on the notion of *influence*—an often-asserted but rarely-interrogated phenomenon in the practice of intellectual history—with an eye towards identifying what historians *mean* when they posit that one event or thinker influenced another. Over the course of this survey, by pinning down this meaning as precisely as possible, we derive a coherent theory of influence which can thus serve as our lens

[41]That is to say, *contextual* rather than textual analyses. If one was solely interested, for example, in evaluating the internal coherence of a single text like *Leviathan*, the synchronic methods already introduced would be sufficient, as we saw in the *Leviathan* embedding examples of Section **??**.

through which to view and understand changes in embedding spaces over time.

### 3.6 Putting it All Together: Networks of Semantic Influence

Although in previous sections we discussed how embedding spaces allow us to visualize the lexical-semantic relationships among individual terms, sentences, documents, and authors, the algorithms for constructing them operate only at the level of individual terms and sentences, such that information is lost when the resulting term or sentence embeddings are coarse-grained up to the document or authorial level. While we are many years away from embedding algorithms that can directly incorporate and model semantics above the sentence level[42], we are able build on recent work in cultural analytics which combines the lexical-semantic modeling power of embeddings with explicit models of authorial influence, allowing us to perform more sophisticated author-level comparisons than would be possible through (e.g.) comparison of mean-pooled sentence embeddings.

In particular, we adopt the approach described in Soni *et al.* 2021 to construct a set of Semantic Leadership Networks which directly model dyadic-temporal influence among major sources of nineteenth-century European socialist thought. For example, to analyze the impact of pre-1848 socialist thought on that of the post-1848 era, we can consider the network in Figure 24, where each edge represents the hypothesized lexical-semantic influence of one node before 1848 on the other after 1848.

Then to estimate the weights on these edges, i.e., the degree of lexical-semantic influence between a dyad of authors, we start by computing the "lead" of one author $a_1$ over another $a_2$ with respect to the movement of their respective embeddings for a particular word $w$ via the following equation, a simplified version of equation (4) from Soni *et al.* 2021:

$$\mathsf{Lead}_{a_1 \to a_2}(w) = \frac{\mathbf{pre}_w^{a_1} \cdot \mathbf{post}_w^{a_2}}{\mathbf{pre}_w^{a_2} \cdot \mathbf{post}_w^{a_2}},$$

[42] Even contextual sentence-level embedding algorithms are still in their infancy: while word embeddings have been in wide use since 2014 (after Mikolov *et al.* 2013, the canonical citation for word embedding-based works, showed their immense effectiveness on a wide range of NLP tasks), with libraries for their construction available in nearly all programming languages by 2016, the first contextual sentence-level embedding library was not made available until 2020 (a year after the canonical publication, Reimers and Gurevych 2019), and at the time of writing no libraries are available for languages besides Python.

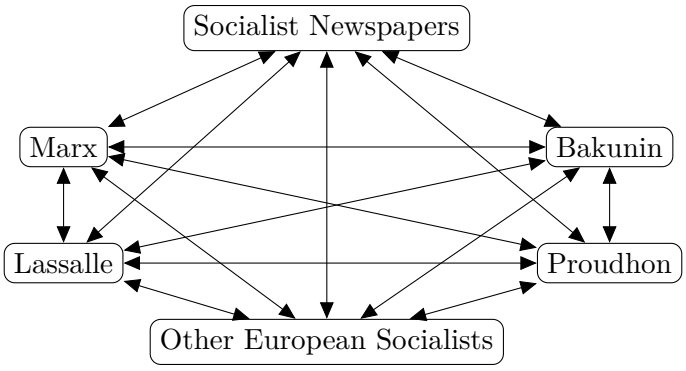



Figure 24: A PGM where we've observed the person's action ($a = \mathsf{Go}$) but only have a probability distribution over the weather $w$.

where the numerator represents the similarity between $a_1$'s pre-1848 embedding for $w$ and $a_2$'s post-1848 embedding for $w$ (both normalized to be of length one, an assumption we relax for the following definitions), and the denominator represents the similarity between $a_2$'s pre- and post-1848 embeddings for $w$. In other words, considering our word similarity function $\mathsf{sim}(\cdot, \cdot)$ described in the previous section, we compute

$$\mathsf{Lead}_{a_1 \to a_2}(w) = \frac{\mathsf{sim}(\mathbf{pre}_w^{a_1}, \mathbf{post}_w^{a_2})}{\mathsf{sim}(\mathbf{pre}_w^{a_2}, \mathbf{post}_w^{a_2})}.$$

As outlined in more detail in Soni *et al.* 2021, this quantity operationalizes our concept of author $a_1$'s influence on author $a_2$ over a given time span: it increases in proportion to how much $a_2$'s post-1848 usage of $w$ is "pulled" towards $a_1$'s pre-1848 usage. For example, Figure 25 visualizes a situation where $a_2$'s post-1848 usage of a word $w$ is drawn away from their pre-1848 usage and towards $a_1$'s pre-1848 usage: intuitively, this can be thought of as $a_2$ being "convinced" to use $w$ in the manner they observed $a_1$ using it, thus dropping their original pre-1848 conception of $w$ over time.

Mathematically, assuming our embedding vectors are all in $\mathbb{R}^{[0,1]\times[0,1]}$, we can obtain a similarity metric $\mathsf{sim}$ that scales from 0 to 1 via the transformation

$$\mathsf{sim}(\mathbf{v}_1, \mathbf{v}_2) = 1 - \frac{\|\mathbf{v}_2 - \mathbf{v}_1\|}{\sqrt{2}},$$

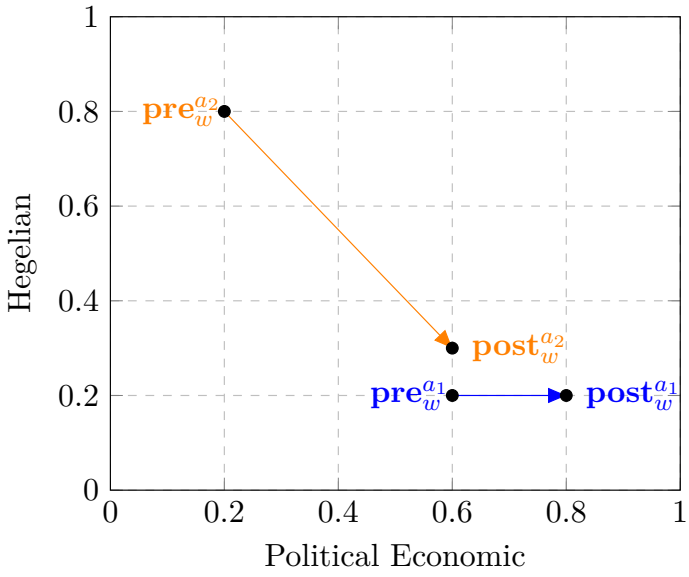



Figure 25: The situation of one-way influence, where $\mathsf{Lead}_{a_1 \to a_2}(w) > \mathsf{Lead}_{a_2 \to a_1}(w)$, allowing us to infer that $a_1$'s usage of $w$ before 1848 had more influence on $a_2$'s usage over time than vice-versa.

where $\|\mathbf{v}_2 - \mathbf{v}_1\|$ represents the Euclidean distance between the embedding vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, and where $\sqrt{2}$ represents the maximum Euclidean distance possible between two points in $\mathbb{R}^{[0,1]\times[0,1]}$. Then we can compute the Euclidean distances as:

$$\begin{aligned}
\|\mathbf{post}_w^{a_2} - \mathbf{pre}_w^{a_1}\| &= \sqrt{(0.6-0.6)^2 + (0.3-0.2)^2} = \sqrt{0.1^2} = 0.1, \\
\|\mathbf{post}_w^{a_2} - \mathbf{pre}_w^{a_2}\| &= \sqrt{(0.6-0.2)^2 + (0.3-0.8)^2} = \sqrt{(0.4)^2 + (0.5)^2} = \sqrt{0.36} = 0.6, \\
\|\mathbf{post}_w^{a_1} - \mathbf{pre}_w^{a_2}\| &= \sqrt{(0.8-0.2)^2 + (0.2-0.8)^2} = \sqrt{0.6^2 + (-0.6)^2} = \sqrt{0.36 + 0.36} \\
&= \sqrt{0.72} \approx 0.849, \\
\|\mathbf{post}_w^{a_1} - \mathbf{pre}_w^{a_1}\| &= \sqrt{(0.8-0.6)^2 + (0.2-0.2)^2} = \sqrt{0.2^2} = 0.2,
\end{aligned}$$

and thus compute our leadership scores for this situation as:

$$\begin{aligned}
\mathsf{Lead}_{a_1 \to a_2}(w) &= \frac{\mathsf{sim}(\mathbf{pre}_w^{a_1}, \mathbf{post}_w^{a_2})}{\mathsf{sim}(\mathbf{pre}_w^{a_2}, \mathbf{post}_w^{a_2})} = \frac{1 - \frac{\|\mathbf{post}_w^{a_2} - \mathbf{pre}_w^{a_1}\|}{\sqrt{2}}}{1 - \frac{\|\mathbf{post}_w^{a_2} - \mathbf{pre}_w^{a_2}\|}{\sqrt{2}}} = \frac{1 - \frac{0.1}{\sqrt{2}}}{1 - \frac{0.6}{\sqrt{2}}} \approx \frac{0.929}{0.576} \approx 1.613, \\
\mathsf{Lead}_{a_2 \to a_1}(w) &= \frac{\mathsf{sim}(\mathbf{pre}_w^{a_2}, \mathbf{post}_w^{a_1})}{\mathsf{sim}(\mathbf{pre}_w^{a_1}, \mathbf{post}_w^{a_1})} = \frac{1 - \frac{\|\mathbf{post}_w^{a_1} - \mathbf{pre}_w^{a_2}\|}{\sqrt{2}}}{1 - \frac{\|\mathbf{post}_w^{a_1} - \mathbf{pre}_w^{a_1}\|}{\sqrt{2}}} = \frac{1 - \frac{0.849}{\sqrt{2}}}{1 - \frac{0.2}{\sqrt{2}}} \approx \frac{0.340}{0.859} \approx 0.400,
\end{aligned}$$

so that we obtain the result that $a_1$'s influence on $a_2$ in this case was about 4 times stronger than $a_2$'s influence on $a_1$.

Importantly, however, the Lead function also allows us to identify instances of *mutual* influence between $a_1$ and $a_2$. In the most extreme case, visualized in Figure 26, $a_1$'s usage of $w$ is influenced by $a_2$'s usage to such an extent that the former's post-1848 usage exactly matches the latter's pre-1848 usage and, with maximal influence also flowing from $a_1$ to $a_2$, $a_2$'s post-1848 usage exactly matches $a_1$'s pre-1848 usage. In other words, $a_1$ and $a_2$ have switched places—both authors drop their pre-1848 usage of $w$ and instead adopt the other author's usage.

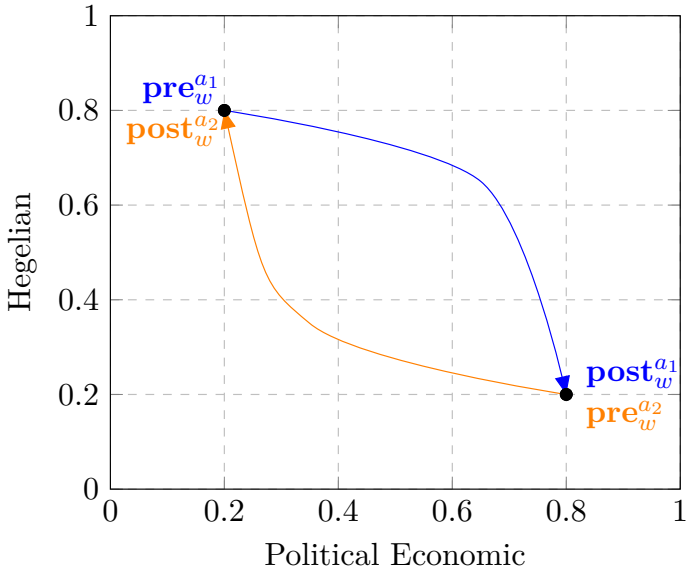



Figure 26: The situation of mutual influence, where $\mathsf{Lead}_{a_1 \to a_2}(w) = \mathsf{Lead}_{a_2 \to a_1}(w)$, since $a_1$'s post-1848 usage of $w$ exactly matches $a_2$'s pre-1848 usage, and $a_2$'s post-1848 usage exactly matches $a_1$'s pre-1848 usage.

In this case, given the values shown in Figure 26, we can compute our Euclidean distances as

$$\|\mathbf{post}_w^{a_2} - \mathbf{pre}_w^{a_1}\| = \sqrt{(0.2-0.2)^2 + (0.8-0.8)^2} = 0,$$
$$\|\mathbf{post}_w^{a_2} - \mathbf{pre}_w^{a_2}\| = \sqrt{(0.2-0.8)^2 + (0.8-0.2)^2} = \sqrt{(-0.6)^2 + 0.6^2} = \sqrt{0.72} \approx 0.849,$$
$$\|\mathbf{post}_w^{a_1} - \mathbf{pre}_w^{a_2}\| = \sqrt{(0.8-0.8)^2 + (0.2-0.2)^2} = 0,$$
$$\|\mathbf{post}_w^{a_1} - \mathbf{pre}_w^{a_1}\| = \sqrt{(0.8-0.2)^2 + (0.2-0.8)^2} = \sqrt{0.6^2 + (-0.6)^2} = \sqrt{0.72} \approx 0.849,$$

so that our leadership scores in this case are:

$$\mathsf{Lead}_{a_1 \to a_2}(w) = \frac{\mathsf{sim}(\mathbf{pre}_w^{a_1}, \mathbf{post}_w^{a_2})}{\mathsf{sim}(\mathbf{pre}_w^{a_2}, \mathbf{post}_w^{a_2})} = \frac{1 - \frac{\|\mathbf{post}_w^{a_2} - \mathbf{pre}_w^{a_1}\|}{\sqrt{2}}}{1 - \frac{\|\mathbf{post}_w^{a_2} - \mathbf{pre}_w^{a_2}\|}{\sqrt{2}}} \approx \frac{1 - \frac{0}{\sqrt{2}}}{1 - \frac{0.849}{\sqrt{2}}} \approx \frac{1}{0.400} = 2.5,$$

$$\mathsf{Lead}_{a_2 \to a_1}(w) = \frac{\mathsf{sim}(\mathbf{pre}_w^{a_2}, \mathbf{post}_w^{a_1})}{\mathsf{sim}(\mathbf{pre}_w^{a_1}, \mathbf{post}_w^{a_1})} = \frac{1 - \frac{\|\mathbf{post}_w^{a_1} - \mathbf{pre}_w^{a_2}\|}{\sqrt{2}}}{1 - \frac{\|\mathbf{post}_w^{a_1} - \mathbf{pre}_w^{a_1}\|}{\sqrt{2}}} \approx \frac{1 - \frac{0}{\sqrt{2}}}{1 - \frac{0.849}{\sqrt{2}}} \approx \frac{1}{0.400} = 2.5,$$

and thus we see that indeed the leadership score is able to capture the perfect mutual influence between the two authors in this case.

Given this intuition around the leadership score with respect to a *word*, the process of aggregating these scores up to the level of documents and authors becomes straightforward: we take, for example, the network from Figure 24 above and assign edge weights such that the weight on the edge from $a_1$ to $a_2$ is the number of times $a_1$ led $a_2$, normalized by the total number of times $a_1$ led *any* author. Mathematically, we have

$$\mathsf{Weight}(a_1 \to a_2) = \frac{\sum_{w \in \mathcal{V}} \mathbb{1}[\mathsf{Lead}_{a_1 \to a_2}(w) > \mathsf{Lead}_{a_2 \to a_1}(w)]}{\sum_{a \in \mathcal{A} \setminus \{a_1\}} \sum_{w \in \mathcal{V}} \mathbb{1}[\mathsf{Lead}_{a_1 \to a}(w) > \mathsf{Lead}_{a \to a_1}(w)]},$$

where $\mathcal{V}$ represents our vocabulary, i.e., the set of all words in the corpus, $\mathcal{A}$ represents the set of all authors in the corpus (so that $\mathcal{A} \setminus \{a_1\}$ represents the set of all authors in the corpus besides $a_1$), and $\mathbb{1}[P]$ is an indicator function, equal to 1 when the predicate $P$ is true and equal to 0 otherwise.

Once we have these weights assigned, the network analysis literature provides several algorithms for computing the influence of one node over others in the network. In our case, following Soni *et al.* 2021, we use the Pagerank algorithm, so that the set of influence scores is the solution to the system of linear equations defined by

$$\mathsf{Influence}(a_i) = \alpha \sum_{a_j \in \mathcal{A} \setminus \{a_i\}} \mathsf{Weight}(a_i \to a_j) \cdot \mathsf{Influence}(a_j) + \beta,$$

where $\alpha$ and $\beta$ are set to be 0.85 and $0.15/|S|$ respectively, and $|S|$ represents the number of authors in the network, following Soni *et al.* 2021.

With this algorithm for constructing Semantic Leadership Networks we thus have a powerful

tool, rooted in linguistically-principled microfoundations, for measuring the relative influences of the authors in our corpus of 19th-century socialist thinkers. In the next two sections we introduce this corpus itself, arguing that it captures a wide swath of 19th century socialist discourse and thus is sufficiently rich for our methods to infer characteristics of this discourse as a whole.

# 4 The Empirics of Influence: Historical Sketches

## 4.1 Theories of Influence, Past and Present

### 4.1.1 Structure vs. Agency

Within the community of intellectual history and the history of political thought, strict battle lines are often drawn between those who center individual autonomy and free will as the driving forces behind intellectual innovation, and those who instead focus on factors outside of the individual such as material conditions, social institutions, or group psychology. Grouping the latter categories under the rubric of "structures", for example, Skinner sees this as a growing concern among post-WWII scholars, to the point that "the question of the relative weight to be assigned to agents and structures has indeed become a central preoccupation with the younger generation of social theorists" (Skinner 1990, 18).

Perhaps especially in debates regarding our topics of study in the chapters that follow, however—Marxism and 19th century socialist discourse more broadly—the adoption or dismissal of key arguments often does hinge upon whether or not one accepts the inviolability of methodological individualism[43].

This general anxiety regarding the importance of methodological individualism manifests, among intellectual historians, as an anxiety around whether historical explanations which employ influence claims strip historical thinkers of their agency. In other words, the main challenge for a theorist in either case is to find a conception of social phenomena which can mitigate the incongruity between (a) looking out at the world and perceiving emergent "social forces" which seem to act on individuals as *objects* despite these forces not having agency or intentionality themselves, and (b)

[43]For the case of debates among late-twentieth-century Western Marxists, for example, see Elster 1982, Althusser 1968, Cohen 1978.

phenomenologically experiencing "selfness" as a subject with the free will to act or not act upon the objects "out there".

The history of Marxism for example, from Marx's time to the present, is largely a history of theorists trying to grapple with this antinomy: if the internal logic of capitalism implies its inevitable self-destruction and replacement by socialism, and with its trajectory towards this outcome determined by iron-clad "laws of motion" discovered by Marx, what exactly is left for an individual socialist to do, free-will-wise (especially given the near-certainty of a violent response by the agents of capital)? The complacency which this theory can therefore induce, invoking images of socialists calmly sitting and waiting for capitalism to fall, has been adduced to explain all manner of failures of Marxian socialism, but especially the failure of the German SPD to seize power in the years leading up to the Reichstag Fire.

Returning to the role of theory in the social sciences, in a paradigmatic case of "easier to be a critic than a creator", much ink has been spilled lambasting social theories on the basis of both (a)—i.e., for granting *too much* autonomy to individuals and not acknowledging the pressures exerted by external forces (for example in critiques of libertarian contract theories)—and (b)—for denying the agency of the individual and treating them as a pawn in the machinations of an external or abstract entity (for example in critiques of "vulgar" or strongly-teleological historical materialism). In our view, however, these argumentative poles can serve as helpful boundary points delineating the two extremes of a spectrum of theories of intellectual influence. The problem outlined at the end of the previous section, of establishing a cogent theory of influence to serve as the interpretive machinery for our diachronic embeddings, thus becomes more manageable; we can evaluate theories on the basis of how well they adjudicate between these two critiques.

In fact, although Skinner frames it as somewhat of a zero-sum conflict in the quotation cited above, Giddens 1979 analyzes aspects of the two "sides" that, we argue, can be made complementary rather than contradictory. He points out, for example, how separate literatures have already made great strides in understanding social behavior at various levels by adopting one or the other approach. Social scientists, he argues, often err on the side of determinism (with an extreme case being, e.g., vulgar historical materialism), while philosophers err instead on the side of theorizing

individual action without reference to social-relational considerations:

> "There exists a large philosophical literature to do with purposes, reasons and motives of action; but it has to date made little impact upon the social sciences. In some part this is understandable, because the philosophy of action, as developed by British and American philosophers, has not paid much attention to issues that are central to social science: issues of institutional analysis, power and social change." (Giddens 1979, p. 2)

In our reading of these facts, then, the historian's task is not to choose between the two, or to cleanly reconcile them, but rather to adopt the contributions of *both* literatures as part of their historiographic toolbox. Hence, combining this interpretation of Giddens 1979 with the later argument of Sperber 1996, we see no necessary contradiction in adopting a notion of intellectual influence that remains committed to materialist and methodological-individualist principles. Consider, for example, the following three points on an Ockham's-razor-based spectrum of theoretical complexity (i.e., a spectrum constructed via a notion of parsimony with respect to the theories' underlying assumptions):

1. *Physical* theories: for example particle physics, which studies the smallest indivisible units known to us (quarks),

2. *Social* theories: for example psychology, which studies the individual as a single cohesive unit, and

3. *Metaphysical* theories: for example Hegel's "world spirit" or the Deep-Ecological Gaia Hypothesis, which assert non-observable agents whose dynamics give rise to observable phenomena.

Considering the points themselves, we can gauge that (1) satisfies methodological individualism trivially (since particles aren't thought of as having agency with respect to their behavior) and (2) satisfies it non-trivially, but that (3) may not satisfy it unless it can be decomposed into combinations of (1) and (2). Then, we can consider the spaces in between the points, and posit e.g. that without any "objective" way to adjudicate among theories lying between (2) and (3), as a

rule of thumb we'd prefer the theory closer to (2). In the next section, then, we choose from among various theories of influence in a similar way, evaluating where they lie on a spectrum of *strictness* with respect to the scope of admissible influence claims, subject to the constraint of not violating key historiographic prerogatives like that of respecting individual agency.

### 4.1.2 Mapping and Evaluating the Theories

Though even the earliest historians implicitly grappled with the complexities of influence claims, it wasn't until the 20th century that historians began explicitly constructing and debating theories of intellectual influence. Assessing the state of the historical profession in his 1926–1928 lectures, for example, R. G. Collingwood laments that historiographic methodology "is in point of fact almost wholly neglected by historians," to the extent that "the ordinary historian can give no account of the processes by which he extracts narrative from sources" (Collingwood 1946, p. 389).

While one can trace this historiographic self-consciousness back to e.g. the inquiries of the Vienna School of logical positivism, the theories most relevant to our investigation were formed in the aftermath of Karl Popper's critiques of theretofore-unexamined aspects of historical practice in *The Open Society and its Enemies* (Popper 1945). Popper massively raised the stakes of the debate around historical influence by arguing that a set of historiographic precepts, introduced by Plato but taken to extremes by Rousseau and Marx, in fact provided crucial ideological support to the self-legitimation of totalitarian regimes. Though Popper produced and presented these arguments, for the most part, in the context of pre-war debates within the philosophy of science, it quickly spread to other fields in the aftermath of World War II as scholars grappled with the horrors that the war had wrought.

Variations on this argument were introduced into political theory proper, for example, by Hannah Arendt's influential *The Origins of Totalitarianism* (Arendt 1951), while J. L. Talmon's *The Origins of Totalitarian Democracy* (Talmon 1952) centered Karl Marx's thought as the apotheosis of this immanently-totalitarian historiographic trend. This illocutionary move to link Karl Marx's thought as inexorably linked with the Soviet regime gained traction in less academic circles as well, as the Cold War heated up in the years following the Berlin Airlift. Prominent US politicians

were spurned into anti-communist action by Arthur Schlesinger's simplified version in *The Vital Center* (Schlesinger 1949), for example, and perhaps most influential in this genre was Theodor Adorno's co-authored *The Authoritarian Personality* (Adorno *et al.* 1950), which gave this theory a micro-level social-psychological interpretation.

We describe this post-war background so as to situate the emergence of the Cambridge School within this broad atmosphere of suspicion and distrust attached to the perpetuation—and even more so the creation—of sweeping "grand theories" of historical development. It is in light of this environment that one can understand the skepticism embodied in the nearly impossible-to-satisfy criteria for valid influence claims developed by Cambridge School practitioners, which we argue are far too restrictive to be fruitfully employed towards interpreting diachronic embeddings. Though their strictness was sometimes attenuated (but sometimes increased) in later methodological works, here we consider the criteria given in Quentin Skinner's "Meaning and Understanding in the History of Ideas" as exemplary of the broader Cambridge School conceptualization of intellectual influence.

In this work, Skinner straightforwardly defines the necessary conditions for a "minimally" admissible[44] influence claim of the form "$A$ influenced $B$":

(i) $B$ is known to have studied $A$'s works;

(ii) $B$ could not have found the relevant doctrines in any writer other than A; and

(iii) $B$ could not have arrived at the relevant doctrines independently.

Our argument that these conditions are excessively strict, however, is based not so much in the content of the conditions themselves than in the epistemic frame in which their necessity is posited. Within an inferential system where statements are evaluated via laws of non-probabilistic predicate logic, like that of the early logical positivists from whom Skinner draws much of his inspiration[45],

[44]Though some discussions of these criteria use "acceptable" or "valid" rather than "admissible" to describe them, the latter two terms introduce the assumption that the theory-maker would believe the influence claim was *true* if it met their criteria, an assumption avoided in the use of "admissible". Although this assumption is probably the case for Skinner, because his conditions are so strict, most theory-makers seem to be outlining something more subtle: conditions that, if they are *not* met, are not "eligible" for true/false inference in the first place. Hence we use "admissible" vs. "non-admissible" throughout.

[45]Here we have in mind, in particular, stronger forms of early logical positivism which contended that statements are meaningful if and only if they can be converted into statements of predicate logic, but allowed for the possibility that meaningful statements whose predicates require sense data for their evaluation (namely *synthetic*, as opposed to *analytic*, statements) may be indeterminate in practice.

it may indeed be the case that the influence claim cannot be evaluated as true unless all three conditions are satisfied. Using such a system to judge a statement like "$A$ influenced $B$" as made by a historian, however, makes sense only under a draconian epistemic regime in which "truth" with respect to historical claims denotes "certainty that all other potential explanations can be ruled out", such that any status short of this leaves the claim consigned to a dustbin of historiography, on the basis that "we'll never know".

Granting here that we indeed will never *know* the truth or falsity of an influence claim in the sense implied this framework, however, we argue that (short of abandoning the historical endeavor altogether) a feasible framework for rigorous inference from historical evidence is still readily available in the form of the standard Bayesian inferential calculus. This approach, shedding the methodological shackles imposed by the need to *prove* truth or falsity of a historical claim, instead allows for varying *degrees of belief* in the truth of a proposition which are—most importantly (in terms of the analytical grounding of our argument)—governed by an "optimal"[46] set of inference rules for updating these degrees of belief upon observing new information. The move from the framework implied by Skinner's criteria to our proposed Bayesian framework can be summarized as changing the evaluative question the historian asks, away from "Is proposition $P$ true?" and towards "How strongly should we believe proposition $P$ to be true?" (Williamson 2010)—in other words, away from yes-or-no questions and towards questions of degree.

Our proposal for adopting this degrees-of-belief approach is motivated by the observation that many of the methods and presuppositions which social theorists import from statistics—having been employed far more frequently and for far longer by the natural sciences—are not up to the task of explaining *social* behavior as-is, i.e., that they will require significant revision rather than cut-and-paste transposition from the natural to the social worlds. This point is argued forcefully, for example, in William Connolly's *The Terms of Political Discourse*, which explicitly argues, "it is unlikely that models of explanation derived from natural science can apply to social and political life in unrevised form" (Connolly 1974, p. 5) As a heuristic for modifying these explanatory models

[46] "Optimal" with respect to a collection of intuitive inferential axioms developed by statisticians, mathematicians, and philosophers in the centuries since the posthumous publication of English minister Thomas Bayes' "An Essay Towards Solving a Problem in the Doctrine of Chances" (Bayes 1763).

to take account of social dynamics (drawing on the pragmatist tradition), Sokal and Bricmont 1997 suggests that we look for inspiration in the ways humans intuitively draw quite complex inferences every day, often without being conscious of doing so. The Bayesian approach meets this challenge: for reasons outside of the scope of this work, this model better captures the way humans neurologically process and store information, and use this stored information to improve future inferences. In evaluating causal questions like "Does cilantro give me a stomachache?", for example, human cognition exhibits the degrees-of-belief approach of Bayesianism rather than the yes-or-no approach of propositional-logic-based models like that employed in Skinner 1969.

While some historians find such an approach unsatisfactory on account of its dependence on "subjective" probabilities rather than objective truths, we argue that this is in fact a point of strength, not weakness, for the veracity of the approach. With respect to its statistical foundations (i.e., when considered as a tool for information processing in general), Bayesian inference is "no more inherently subjective than are likelihoods and the repeat sampling assumptions required [by non-Bayesian approaches] for significance testing" (McElreath 2020, 35). Thus, we argue, when one takes it and applies it to historical phenomena, its strength comes not from the fact that it eliminates uncertainty (no inferential procedure does) but from the fact that it does the exact opposite: that it explicitly *includes* subjectivity, uncertainty, and ambiguity as considerations which the model itself can reason about. "Because the prior [the representation within a Bayesian model of the researcher's background assumptions] is an assumption, it should be interrogated like other assumptions: by altering it and checking how sensitive inference is to the assumption." (McElreath 2020) By way of this sensitivity-checking process, moreover, historical researchers can explicitly test the robustness of their findings. For example, among classical scholars, Plato's *Epistles* are "among the most disputed texts of antiquity" (Wohl 1998), with scholarly views regarding which of the 13 proposed letters are authentic spanning from "none" to "all". Thus, if a historian who adopts a Bayesian framework is concerned about whether their findings depend too strongly on the contents of some or all of these letters, they can simply remove the questionable entries from the set of admissible evidence encoded in their model's prior, conduct the inference via the same procedure as before, and see whether or not their previously-obtained results remain significant.

If one abstracts away from the particular Bayesian terminology of this conception, moreover, it simply describes an approach to historical understanding which was consciously articulated well before the advent of the Cambridge School. Collingwood's 1926–1928 lectures for example, mentioned at the beginning of this section, argue that "every narrative that we can at any given moment put forward is only an interim report," but that, importantly, "it does not not follow that there can be no solid advance in historical knowledge" (Collingwood 1946). The Bayesian machinery, we argue, simply adds flesh to the skeleton of this insight, allowing us to reason in a principled manner about what the most *likely* narrative(s) are with respect to the currently-available evidence. Indeed, if we substitute the phrase "most likely" for the term "true", this approach satisfies the criteria Collingwood proposes as the requirement for such "solid advance[s] in historical knowledge" to be made: that "the principles [of historical inference] must be independently established *a priori* in order that the narrative constructed by their means may be known to be true" (*ibid.*, 389). This move away from binary evaluation and towards a framework allowing for a continuum of belief, therefore, better reflects the hermeneutic ideals of historians and readers of history and gives us a richer procedure for evaluating influence claims than the "true only if no other possible explanation" framework behind Skinner's conditions.

It is worth noting, however, that although post-WWII skepticism led to their excessive strictness, these criteria did serve as a much-needed corrective to a historiographic paradigm that was only beginning to fall out of favor in 1969 (38 years after it was named and critiqued by Herbert Butterfield): the paradigm of "Whiggish" teleological explanation. In short, works in this paradigm typically construct a historical narrative wherein a well-ordered sequence of "inevitable" events culminates in the form of an event the historian sought to explain, which thus serves as the narrative's climax. Examples abound, but Hegel's presentation of constitutional monarchy as the completion of a world-historical progression, or modern hagiographies of US athletes and entrepreneurs valiantly pushing themselves beyond all obstacles to their success, serve as examples of this tendency.

## 4.2 The Empirics of Influence: Historical Sketches

Having now introduced the models of inference we will employ, in this section we move from the theory to the practice of historical inference. In a series of case studies, we develop intuition around how the qualitative, interpretive claims made by political theorists can be translated into quantitative models, thus building towards the more in-depth studies we conduct in the chapters that follow.

If we hope to expand the horizons of historical research by bringing the power of well-established quantitative methods to bear on our understanding of political texts, we first need to understand how exactly these texts can be brought from the physical into the digital realm, and get a sense for the landscape of already-digitized texts which can thus immediately be used as inputs to our computational analyses.

Luckily for us, vast collections of historical texts and historically-relevant data have been digitized in recent years, such as the millions of books made available by the HathiTrust Partnership[47]—providing resources for which (we argue) supply currently far outpaces demand. A computational study of influence in intellectual history could leverage these resources to develop a set of initial hypotheses regarding who influenced whom—and what types of mechanisms or relationships facilitated this influence—when could then be operationalized as inputs to the embedding methods described above and evaluated based on their results.

In fact, we contend that many of the historical claims made in the Cambridge School works discussed above suffer—despite their invaluable incorporation of contextual information in the form of other *texts*—from a lack of attention to the *non*-textual metadata available about the texts and authors under consideration. To take one prominent example, consider Skinner's analysis of E. M. Forster's *A Passage to India* in the first volume of *Visions of Politics*. A seemingly insignificant part of Forster's novel, its concluding sign-off—"Weybridge, 1924"—is considered in the context of literary *conventions* around sign-offs: James' Joyce's "Trieste-Zürich-Paris" sign-off at the end of *Ulysses*, for example, is taken by Skinner as an example of the sign-off's role in "draw[ing] attention to the romantically nomadic life of the author" (123). Thus, by signing *A Passage to India* off with

[47]In fact, the books in the HathiTrust library are not only digitized but also preprocessed via natural language processing algorithms, enabling their use as direct inputs to many of the algorithms described and used in this work.

the simple reference to Weybridge—"the classic instance of a prosaic English suburb" in Skinner's words (124)—Forster is seen as *doing something with words*, namely, subtly mocking the "posturing" (*ibid.*) of those who use this authorial convention to portray themselves as cosmopolitan, globe-trotting intellectuals, traveling the globe to give a glimpse of its splendor to the unwashed masses.

Bevir 1999 astutely points out, however, the fact that *textual* evidence on its own, from Forster's novel itself or the broader context of the sign-offs of contemporary novels, is not sufficient to justify Skinner's conclusion. Skinner may be right, but it also may simply be that "he did write the novel in Weybridge in 1924, and maybe he intended simply to record this fact." Thus, to decide between these two hypotheses, "we must consider things other than the linguistic context of the novel." (86) Indeed, once equipped with non-linguistic contextual data about Forster's life, adjudicating between the two interpretations becomes a fairly straightforward task: "If they discover that he wrote the novel in Cambridge and India from 1922 to 1924 without ever visiting Weybridge, the case for understanding the utterance as a parody would seem more or less conclusive. But if they discover that he wrote the novel in Weybridge in 1924, the idea that his utterance was a parody would begin to look rather more doubtful." (*ibid.*) It is this example of the insufficiency of purely textual data for contextual history, but the possibility of settling disputes via contextual *metadata*, that motivates our sketch in the next section of bringing Wikipedia data to bear on discussions of influence in the history of political thought.

### 4.3 Text-Mining Influence Claims

Though it has not, to our knowledge, made its way into the academic literature, a series of projects from data-visualization hobbyists over the past few years have shed new light on the "flow" of influence among philosophers by constructing diachronic plots constructed from contextual meta-data contained in Wikipedia pages for philosophers, intellectuals, and political actors, in particular, from the "Influenced By" and "Influenced" sidebars often found on the pages for these individuals. Beginning with a 2012 blog post titled "Graphing the History of Philosophy"[48], the project was expanded to cover the entirety of Wikipedia in 2014[49], and this community has continued to explore

[48] https://coppelia.io/2012/06/graphing-the-history-of-philosophy/

[49] No longer on the web, but available via the Wayback Machine archive.

the possibilities of what can (and can't) be learned from this data, whether automatically[50] or in combination with manual curation[51].

Inspired by this endeavor, and how the developed tools could aid studies of influence in the history of political thought, we began by similarly scraping the data contained in all instances of this sidebar across Wikipedia. Crucially, we were then able to augment this base data with both the automatically-generated and manually-curated data generated by the data-visualization projects described above, resulting in 5,033 hypothesized influencer-influencee pairs[52].

Taking these influencer-influencee pairs as seeds, possibilities abound for "bootstrapping" further pairs by computationally mining actual texts (both pedagogical and scholarly), using the types of language models described earlier to identify:

(a) sentences, paragraphs, chapters, or other textual units[53] where both thinkers in the pair appear, as well as

(b) sentences which mention "influence", "mentorship", or other influence-related words, which may or may not refer to any particular thinker.

As an example of a concrete step in this direction, we obtained JSTOR's data[54] for all issues of the *Journal of the History of Ideas* from 1940 to present and all issues of the journal *History of Political Thought* from 1980 to present[55]. With this data in hand, in a manner akin to Rawls' method of reflective equilibrium[56], we refined the base hypotheses further and further by

1. Analyzing random samples of the type (a), to discover alternative linguistic forms in which influence claims are made, and then

[50] See, e.g., https://github.com/S4N0I/theschoolofathens

[51] https://www.denizcemonduygu.com/philo/

[52] This data is available at https://github.com/jpowerj/our-word-is-our-weapon

[53] For the sake of brevity, but without loss of generality, we restrict our discussion to sentences for the remainder of the section.

[54] Specifically, via a combination of JSTOR's data export tools and our own scraping code, we obtained PDFs, OCRed text, $N$-grams (up to $N = 3$), and metadata (authors, dates, journal and issue names, etc.) for each paper.

[55] Also available at https://github.com/jpowerj/our-word-is-our-weapon

[56] See Rawls 1951, which computer scientists would recognize as a very similar process to that of the Expectation Maximization (EM) Algorithm in Machine Learning (where in the latter instance the hypotheses happen to take the form of numerical estimates of a model's parameters).

2. Analyzing sentences of type (b) to reveal new influencer-influencee pairs that had not previously been considered.

Once this procedure had been carried out to the point that diminishing returns overtook the benefits—in our case, the admittedly subjective point at which an HPT expert could attest that it covered a substantially wide range of the types of influence claims made in the field—the data (i.e., the union of the influencer-influencee pairs themselves and the sentences in our corpora which pertained to them) was used as the input to two separate machine learning models. In the first, a binary classifier was trained to differentiate between influence claim statements and all other statements, and (importantly) to provide confidence scores for each classification, such that we were able to add additional statements to our collection of influence claims by running the trained model on the full HPT journal corpus and extracting claims classified above a minimum-confidence threshold as influence statements.

Then, to take a step from simply *collecting* these influence-statements towards *understanding* them, we trained a more complex multi-class classification model to *categorize* the influence claims based on the mechanism(s) of influence they posit, from among the mechanisms discovered via step (b) in the procedure discussed above. In this case, the model was trained to differentiate between claims based on:

1. Collaboration: for example, "Karl Marx and Friedrich Engels carried on one of the great intellectual collaborations in the history of scientific research." (Gandy 1979, p. ii)
   $\Longrightarrow$ Collaboration(`Karl Marx`, `Friedrich Engels`)

2. Mentorship[57]: for example, "the grandly-titled *Tractatus Logico-Philosophicus* was begun in 1915 following Wittgenstein's tenure at Cambridge under [Bertrand] Russell's mentorship" (Wright 2006, p. 72)
   $\Longrightarrow$ Mentorship(`Bertrand Russell`, `Ludwig Wittgenstein`)

[57] A recent quantitative work, published after the conclusion of our small-scale study, was able to scrape 37,157 mentor-mentee pairs from the ProQuest PhD Dissertation and Thesis databank (PQTD), "an official record of advisor–student relationships taken from PhD theses" in biomedicine, chemistry, math, and physics between 1960 and 2017, "supplemented with crowdsourced data from Academictree.org and the Mathematics Genealogy Project (MGP)." (Ma *et al.* 2020, 14077) The creation of a "ground truth" dataset of this magnitude for the history of political thought (rather than the natural sciences) would, we argue, represent a watershed moment for the empirically-principled study of the role played by influence in the development of political thought.

3. Reading: for example, "Not only had Marx read Hegel's *Logic* in 1858, but we know that he studied it once again in 1860." (Dussel 2002, p. 195)
   $\Longrightarrow$ Reading(`Karl Marx`, `G.W.F. Hegel`)

4. Shared-Intellectual-Community: employment or study at the same institution, for example, as in "Schelling and Hegel first met at the Tübingen Seminary in 1790, and the two young men shared a room there during their student years" (Levine 2006, p. 124)
   $\Longrightarrow$ Shared-Intellectual-Community(`F.W.J. Schelling`, `G.W.F. Hegel`)

5. Shared-Geography: for example, "Lenin was an occasional visitor in Vienna. Hitler, like Trotsky, had lived there for years." (Morton 1990, p. 290)
   $\Longrightarrow$ Shared-Geography(`V.I. Lenin`, `Adolf Hitler`)

6. Rivalry: for example, "the cultural philosophy of the Historical School and, more particularly, of Hegel's rival at the University of Berlin, Friedrich Schleiermacher." (Toews 1985, p. 7)
   $\Longrightarrow$ Rivalry(`G.W.F. Hegel`, `Friedrich Schleiermacher`)

We note here that the goal is not to classify the hypothesized influencer-influencee *pairs* using this schema—this would introduce a new set of complexities due to the confoundedness of our categories (specifically, the presence of one relationship makes the presence of the others much more likely)—but only the *statements* making influence claims. *After* the statements themselves have been classified by our model, we subsequently label the referenced pairs with the relationship type learned for each statement.

We note this detail because it represents a situation which may be counter-intuitive for those unfamiliar with what is called the "bias-variance tradeoff" in statistical inference: for example, although one might have the intuition that the best model would be one able to capture the *multiple* relations posited by a compound statement like "Hegel and Schelling were university roommates, early friends and collaborators, and eventually rivals," in practice to the small-$N$ nature of the study (relative to the datasets with billions of observations regularly used by computer scientists in academic and commercial settings) means that the greater statistical power of a simple model distinguishing between four classes far outweighs any benefits of a more complex Multi-Label

Classification model which, though it *can* handle these compound statements, massively sacrifices statistical power by having to distinguish between $2^4 = 16$ possible combinations of classes for each statement. In fact, in our case, the multi-label classification approach is unnecessary, as even compound statements like these can be decomposed by way of linguistic dependency parses[58].

Then, by examining which linguistic features in the input statements best helped the algorithm distinguish between these classes[59], we can understand in a linguistically-principled sense *how* historians of political thought talk about influence, and (simultaneously) automatically generate detailed timelines of posited influence paths, with each "link" labeled with its mechanism.

As the final step in our derivation of a fruitful model of intellectual influence, we linked these empirical findings with the Bayesian model described earlier, which allowed us to derive measures of uncertainty with respect to each of the findings. Though this could be accomplished in several ways (for example, as in the example of Plato's *Epistles* above, one could reduce the degree of certainty attached to some specific source to 1/2 the level of all other sources), in our case we assigned weights representing our conception of the relative strength with which each influence mechanism acts. For example, to encode the supposition that the strength of the Shared-Geography relation tends to be lower than that of the Collaboration relation[60] in terms of explanatory power (as an anecdotal but certainly noteworthy example, Hitler, Trotsky, Tito, Freud, and Stalin all lived in Vienna in the early 1910s[61], but several pairs within that group were most likely not aware of each other),

[58] See Ash *et al.* 2020 for an example in computational social science, where this approach is used to differentiate and separately model the subjects and objects of legal contract clauses.

[59] The particular type of classifier we trained, called the Random Forests model (as implemented in the scikit-learn library's `RandomForestClassifier` package), produces an explicit decision tree which is constructed to optimally partition—in the sense of minimizing the entropy or Gini coefficients of successive splits—the domain of the inputs given a set of pre-specified constraints like the maximum depth of the desired tree. Hence one can "read off" the most helpful features directly from this tree, as (for example) its topmost decision node will split the data based on the input feature most predictive of the output, the nodes in its second level will split the data based on the input feature those most predictive of the output within the subgroups created by the first split, and so on.

[60] Here we highlight another commonly-misunderstood strength of the Bayesian approach: though its critics within social science and the humanities often decry the presumptuousness of "assigning a number" to an abstract phenomenon like the strength of a type of influence, in fact what matters in nearly all of these models is not the individual *values* of the numbers used to encode prior assumptions, only their relative magnitudes. Mathematically, this fact can be derived in a manner similar to the econometric argument for focusing on ordinal, rather than cardinal, utility in models of welfare economics (see, e.g., Harsanyi 1953 for the basic argument, and Roemer 1996, 13*ff.*, for an example of its relevance for conceptualizing, analyzing, and comparing political-theoretic models of justice such as those of Rawls, Nozick, and G. A. Cohen).

[61] See Morton 1990 (cited above for our example Shared-Geography statement) for details on their meetings and near-meetings during the years 1913 and 1914.

we set the priors in our model such that the effect on the model of observing a Shared-Geography datapoint was scaled to 1/3 of the effect of observing any other type of datapoint[62].

The end result of inducing this prior, then, is that our lower confidence in claims of influence based on Shared-Geography relations translates into a reduced *posterior* degree of belief generated by the model for influence pairs which are heavily based on Shared-Geography claims, relative to the degrees of belief that it would generate for these pairs if all four types of claims were weighted equally. Thus we see that, as discussed above, adopting this Bayesian framework renders our modeling assumptions maximally transparent, to the point that it in fact provides other researchers studying the same phenomena with a set of tools for direct constructive criticism of the model: they can point to particular explicit assumptions in the priors that they believe to be erroneous or invalid, remove them (for example, by weighting all four types of influence claims equally), and see if our findings still hold.

## 4.4 The Point is to Change It: Theoretical Innovation and Political Practice in the History of Marxism

In this section we introduce the basics of how computational text-analysis algorithms can be used to fruitfully study intellectual influence, by analyzing a corpus of prominent Marxist texts from 1848 to the present and exploring what types of influence relations we can capture among its texts. Specifically, we compare findings derived using an information-theoretic method developed in Barron *et al.* 2018 with those of a probabilistic text-analysis method developed in Gerow *et al.* 2018, both of which aim to quantify the *novelty* and *influence* of particular texts within the corpus. We illustrate throughout how these computational tools can be linked with social-scientific theories, by interpreting the algorithms and derived quantities as capturing the creation and destruction of ideological "repertoires of contention" within Marxist discourse, as defined in Tilly 2008.

Using this combination of computational and social-scientific models, we then move to a confirmatory, rather than exploratory, analysis, testing a set of hypotheses regarding the influence rela-

[62]Specifically, in terms of the mechanics of the Bayesian model, this scaling was achieved via Laplace smoothing, which induces a particular prior but can be easily interpreted as telling the model to perform inference as if it had observed each of the *non*-Shared-Geography datapoints three times, but the Shared-Geography datapoints only once.

tionship between texts and the historical contexts in which they were published (in the language of the model, the relationship between the $E_t$ and $X_{i,t+1}$ nodes). We utilize statistical change-point estimation methods (Kulkarni *et al.* 2015) to test whether the Soviet invasion of Hungary in 1956 sparked a statistically-significant shift in the novelty of texts in the Marxist tradition. A significant positive shift would potentially indicate a search for new approaches among Marxists who turned away from "official" Soviet communism as a result of the invasion[63], whereas a significant negative shift could be interpreted as evidence of a successful clampdown on critical discussion of the invasion by the Politburo.

The first step in our study was to develop inclusion criteria for a corpus of Marxist texts sufficiently representative of both strands of thought and time periods in the history of Marxism. While in our full study of the origins and diffusion of Marxism in the coming chapters we focus more on these criteria, for the sake of this research sketch we simply adopt the two-column timeline of major texts and major events in Marxism from McLellan 2007.

With the texts we'd like to include in our corpus now established, the next task was to determine which of them already exist in digital form, and which we would need to digitize ourselves. By searching in HathiTrust, Archive.org, and ProQuest collections, we were able to obtain already-digital versions of 90% of the texts in the corpus, leaving only 6 that we had to digitize ourselves, by scanning the physical books borrowed from the library and then running Optical Character Recognition software on the scanned .pdf files[64].

For the next step, we processed these digitized texts so as to extract from each book the data necessary for input to the algorithms of Barron *et al.* 2018 and Gerow *et al.* 2018. In the former case, the algorithm operates by (a) training a 500-topic LDA topic model on the *full* corpus, without considering the timestamps of the texts, (b) using this corpus-wide topic model to compute topic distributions for each time slice, and then (c) computing information-theoretic measures for pairs of time slices, for example between the topic distributions at times $t$ and $t + 1$.

The results of the study are summarized in Figures 27a and 27b, where each "block" represents

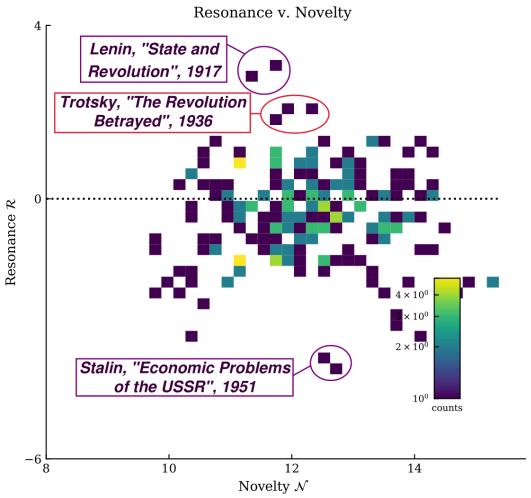
[63]See Hudson 1992, for example, for the finding that "the combined effect of the events of 1956 led to a membership loss in the region of 7,000" for the Communist Party of Great Britain, mainly due to its official support for the invasion.

[64]Specifically, we used Version 15 of Abbyy FineReader Pro for Windows.

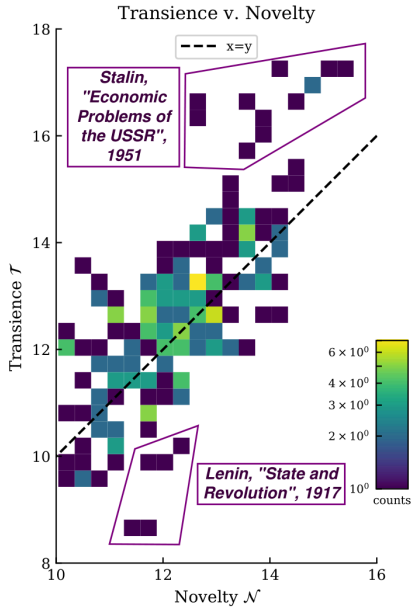
a quantized portion of the novelty-resonance space, so that for example the green block lying on the dashed line in Figure 27b at the point where novelty and resonance both equal 12 represents the fact that four documents had resonance and novelty scores between 12 and 12.333 (the quantization cutoff)[65]. Three noteworthy outliers immediately become apparent: Lenin's *State and Revolution* and Trotsky's *The Revolution Betrayed* stand out as the two highest-resonance works across the corpus, while Stalin's *Economic Problems of the USSR* stands out as the lowest-resonance work, despite all three being comparable in terms of their novelty with respect to previously-published works. The extraordinarily low resonance score for Stalin's last major publication before his death sheds light, for example, on the severe impact the events of 1956 had on subsequent Marxist thought: on top of the previously-discussed effects of the invasion of Hungary on the global communist movement, Khrushchev's "Secret Speech" of the same year—and the newfound possibilities for Communists to criticize Stalin that came as a result—almost instantly destroyed the standing of Stalin's thought within this (admittedly Western-biased) corpus of Marxist literature. To see this, note that Stalin's *earlier* works were among the highest-resonance works in the corpus, with most having a resonance score above zero (i.e., above the dotted line in Figure 27a), with around the same novelty score as *Economic Problems*.

The massive difference in the impact of this latter work (operationalized through its resonance scores), then, can plausibly be ascribed to some combination of his death and the Secret Speech. While more work will need to be done, above and beyond this sketch, to investigate this hypothesis in more detail (for example, if posthumous works of Stalin published *before* the Secret Speech also display this property of extremely low resonance, we have evidence in favor of his death as the more impactful event with respect to the cessation of Stalinist lines of thought among Western Marxists), it hopefully serves as another example of how computational tools can aid—without replacing the role of humans—in text-focused and context-sensitive historical research.

[65] In order to eliminate the potential effects of the texts' *lengths* on their resulting scores, we partitioned each text into equal-length chunks (of approximately twenty pages with respect to a standard printed English translation, so that e.g. Lenin's *State and Revolution* is split into exactly five chunks).




(a) Novelty and resonance scores for each text in our corpus of influential Marxist texts.




(b) Transience and novelty scores for each text in our corpus of influential Marxist texts.

## 5 Conclusion

We began by introducing the hermeneutic background and inferential principles of the Cambridge School of historical analysis, and brought them into conversation with the computer science literature on contextual word embeddings. We showed how the latter not only *implicitly* implements the Cambridge School methodology in a quantitative form but is in fact *explicitly* linked to it through a shared history, which bifurcated into computational-linguistic and linguistic-philosophical strands in the 1950s, losing their shared language to an extent in the intervening years.

After describing these implicit and explicit connections, we dove more deeply into the details of how exactly computational embedding algorithms capture the contextual relationships between words: by constructing a geometric space mapping each word to a point such that points closer together share more contexts, or (in the case of the more recent `BERT`-style models) mapping each word to a *set* of points capturing different senses in which the word is used while still retaining the semantic meaning of distances between points.

We then took a quick detour through structuralist linguistics to explain the fundamental distinction between *synchronic* and *diachronic* modes linguistic analysis, to point out how these algorithms

remain within the synchronic mode (modeling language as a single, static geometric space), and thus need to be paired with a diachronic model to constitute a sufficient tool for *social-scientific* analysis of language. Considering *influence* as the fundamental diachronic relationship of interest to historians of political thought, we reviewed theories and models of intellectual influence with the goal of identifying weaknesses in the standard Cambridge School approach (that is, the approach described in Skinner 1969), and constructed a new model able to overcome these weaknesses.

Next, to preview how this model can be employed towards fruitful analysis of historical texts, we walked through a series of introductory case studies, each one highlighting a particular feature of the model which our studies in the remaining chapters will take advantage of. In the mining-influence-claims case study, for example, we demonstrated a data-driven method for constructing an ontology of influence claims in the format of our model, focusing throughout on how assumptions regarding this ontology (for example, that Collaboration is a stronger influential force than Shared-Geography) can be transparently encoded—and then straightforwardly changed—as parameters within the model.

Finally, in the previous section, we introduced the Semantic Leadership Network model, which explicitly combines the synchronic embedding algorithms described in Section **??** and the diachronic influence model described in Section **??** by constructing *networks* of embeddings where each node represents the embedding of author $a$'s writing at time $t$ and each edge is a connection between a time-$t$ node and a time-$t+1$ node, representing a potential instance of temporal influence. We concluded this section with the final three pieces of the puzzle, namely, (a) the mathematical formulation $\mathsf{Lead}_{a_1 \to a_2}(w)$, which measures author $a_1$'s influence on author $a_2$ by way of their respective embeddings for a word $w$; (b) the method of edge-weighting by which the network of *word*-level $\mathsf{Lead}_{a_1 \to a_2}(w)$ scores is coarse-grained up to the level of author $\times$ time pairs (in our nomenclature, a network of $\mathsf{Lead}^*_{a_1 \to a_2}$ scores); and (c) the use of the PageRank algorithm and its variants to derive our final author-influence scores.

By laying this groundwork—in particular, by walking through how the models emerge naturally from philosophical and social-scientific considerations, not from arrogant pretensions of automating or "solving" the process of historical inquiry—we hope to have preemptively quelled anxieties which

may arise in the remaining chapters regarding what assumptions we're hiding in the machinery of the models, or what types of conclusions we are and are not comfortable drawing from their outputs. If we are successful in this endeavor, we hope in turn to have nudged the discussion around works of computational social science slightly away from these methodological fears and towards methodological possibilities for understanding political thought.

# A Probabilistic Graphical Models in Political Theory

## A.1 General Graphical Models

When trying to understand a complex historical phenomenon one which involves lots of "moving parts"—people, institutions, social relationships, and other related events—interacting to produce the event of interest, researchers are faced with the daunting task of where to begin their investigation. In this study we are indeed confronted with such a challenge, as we aim to understand the dynamics of three decidedly complex historical phenomena:

1. the French Revolution,
2. the intellectual origins of Marxism, and
3. the spread of Marxism across Europe and, later, the Third World.

Just as a mountain climber employs "rules of thumb" in deciding where to start their climb, in our study the following rule of thumb determines the course our investigation: that, to understand a complex historical phenomenon, the most fruitful way to make progress is to (a) break it down into its constituent elements, and then (b) specify how these elements work together to produce the phenomenon. This "decompositional" framework, we argue, corresponds naturally to a particular statistical framework known as Probabilistic Graphical Modeling. As we will argue below, this approach provides a method for transforming our *intuitions* regarding events and their interrelationships into rigorous, measurable, and testable statistical *hypotheses*.

A Probabilistic Graphical Model (PGM) is a statistical tool which, when paired with the laws of statistical inference, formalizes the decompositional framework we outlined above. Concretely,

a PGM is a collection of nodes (drawn as circles), representing variables, and edges (drawn as arrows), representing relationships of influence between nodes—relationships which are then codified numerically in the form of "Conditional Probability Tables". This framework provides a level of abstraction that makes it invaluable for historical research, in that a wide range of historical-analytical frameworks can be subsumed (and thus formalized and tested) within it. Putting the "probabilistic" aspect aside for a moment, we can consider using a graphical model to begin a historical inquiry into the origins and eventual trajectory of the French Revolution, for example:

- "Mainstream" historians can begin modeling the event by considering nodes for the Nobility, the Clergy, and the Third Estate, and edges for their pairwise interactions, as in Figure 28.
- Marxist historians can instead consider the nodes to be the Bourgeoisie, the Proletariat, and the Peasantry, with edges again representing pairwise interactions, as in Figure 29.
- "Great Men" theorists can model prominent individuals as nodes: one for Louis XVI, another for Robespierre, a third for Napoleon, and so on, as in Figure 30. In this case, we see that *not* all pairs of nodes have edges between them, since Louis XVI and Napoleon did not (at least, in a standard interpersonal sense) interact as part of the Revolution.

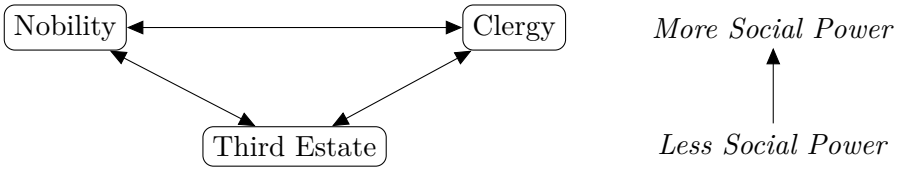



Figure 28: A graphical model illustrating the initial decomposition which a historian of the French Revolution might perform to make their analysis more tractable: three nodes representing three historical entities—the Nobility, the Clergy, and the Third Estate—and edges representing the interrelationships between each pair of entities.

As these figures demonstrate, graphical models provide the researcher with a vast set of modeling possibilities: nodes can represent people, classes, institutions, or other units of observation, while edges can represent interpersonal interactions, economic relations, or any other dyadic relationships between nodes[66]. In addition, given the graphical representation, the *position* of the nodes in 2D

[66]In fact, although the probabilistic aspect of PGMs—the ability to derive statistical inferences regarding the nodes

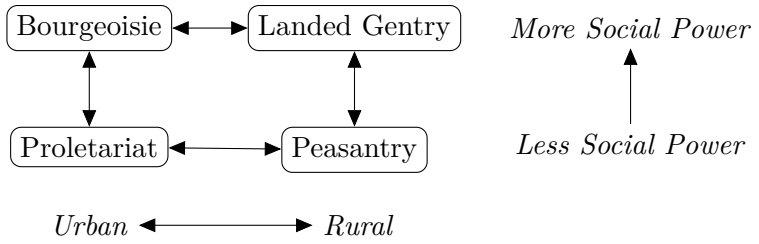



Figure 29: A graphical model illustrating the initial decomposition which a Marxist historian of the French Revolution might perform to make their analysis more tractable: three nodes representing three historical entities—the Bourgeoisie, the Proletariat, and the Peasantry—and edges representing the interrelationships between each pair of entities.

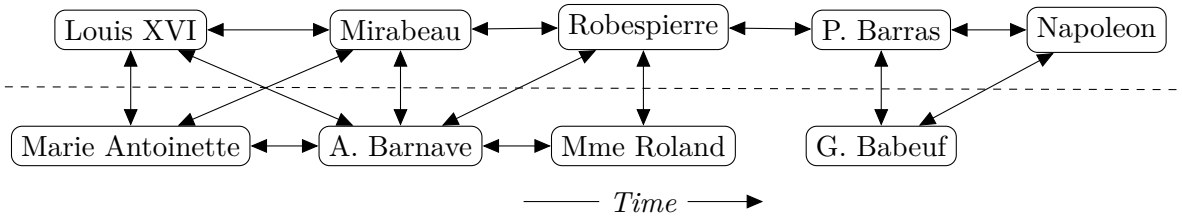
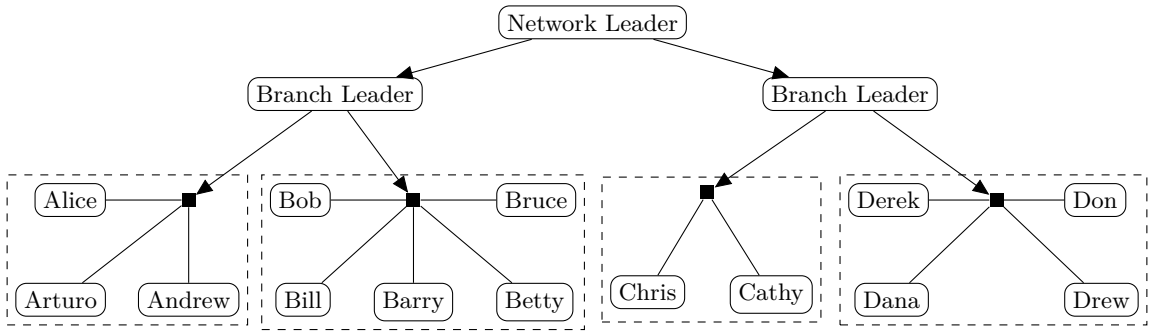



Figure 30: A graphical model illustrating the initial decomposition which a "Great Man" theorist might perform to make their analysis of the French Revolution more tractable, with nodes for prominent individuals and edges between individuals who are known to have interacted.

space can also be used to represent divisions or axes of differentiation: in the case of Figure 29, for example, moving left to right represents moving from urban-situated to rural-situated entities, while moving from bottom to top represents moving in the direction of increased social power. Similarly, in Figure 30, moving left to right represents the passage of time (with nodes positioned

and their interactions—requires edges to be dyadic, for the purposes of non-probabilistic modeling one can introduce "hyperedges" representing a relationship between any number of nodes. The Clandestine Cell Structure of modern insurgent groups, for example, can be modeled effectively using hyperedges, since it is the holistic relationship among *all* members of a cell, rather than the individual dyadic relationships between them, that is most relevant to the functioning of the structure:



based on when, in the course of the revolution, their role was most prominent), and the nodes are divided vertically into two groups based on whether or not the individual was the main political ruler of France at a given time. With only a few nodes this spatial organization may seem like more trouble than it's worth, but it becomes immensely helpful as the model becomes more and more complex. For example, if the Marxist researcher aimed to subsequently model the Soviet subdivision of the Peasantry into *bednyak*, *serednyak*, *kulak*, and *batrak* classes, the *batrak* class could be placed somewhere in the middle of the rural-urban axis, as members of this class often migrated seasonally between urban and rural work.

As these examples show, Graphical Modeling provides a powerful tool for developing explanatory theories of historical phenomena, in forcing the researcher to think simultaneously about (a) what the relevant entities are and (b) what types of relationships exist among them. Thus far, however, we have left out the "probabilistic" aspect of the models used throughout this paper, by remaining agnostic about what particular types of relationships are admissible in these models. Now that we've shown the usefulness of general Graphical Modeling for theory development, in the next section we show—by way of an introduction to text-analytic topic models—how *Probabilistic* Graphical Modeling further enhances the researcher's toolkit, by enabling principled statistical inferences to be drawn regarding the interrelationships between nodes in a Graphical Model.

## A.2 The Role of Probability

While in the previous section we restricted our examples to general historical modeling, in this section we turn to the use of *textual* data in conducting a historical study, and in the process show how the probabilistic aspect of the Probabilistic Graphical Model allows us to reason in a statistically-principled manner about what a given corpus of texts tells us about the historical phenomena we're interested in studying.

The first step in a researcher's journey towards understanding a textual corpus is to develop a rough schema of topics covered in the corpus, to figure out what each text is "about", and to group the texts accordingly. This thematic categorization, the act of transforming an archive of texts into one partitioned into sections, often takes up a massive chunk of research time and resources.

Given this resource bottleneck, this transformation is precisely what one of the first text analysis methods, Latent Semantic Analysis (LSA), was created to do, and (somewhat miraculously) it does so without any input required from the researcher besides having the texts in some digitized format.

Scale-wise, LSA is already leaps and bounds beyond human capabilities, but we can do better: in 2003, David Blei, Andrew Ng, and Michael I. Jordan developed an extension to LSA called Latent Dirichlet Analysis (LDA), which "zooms in" on each document and actually learns distributions over topics for each token (word). That is, while LSA places each document into a single category, LDA derives a more detailed summary of each document, like "25% of this document is about computer science and 75% is about linguistics," a more realistic model given the tendency for most written documents to range across multiple topics (for example, a news article introducing a new technology and then discussing its potential societal impact). In fact, if a researcher does want a single category for each document, they can simply choose the topic with the highest proportion: linguistics, in the case of our example document.

For a researcher hoping to study taxation practices in *ancien régime* France, for example, this means the difference between reading through *every* text in the archive and reading only the specific subset of the texts which are known to discuss the topic of taxation. Depending on the time and resources available to the researcher, for example, they can read the $N$ documents with the highest proportions devoted to taxation: the more resources are available, the higher this $N$ can be.

Both LSA and LDA fall into the category of "unsupervised" algorithms, given the lack of user intervention in the topic-learning process. While this approach stays true to the idea of "pure" exploratory data analysis, it is rare for a researcher to have absolutely no idea what topics lie within an archive. More commonly, researchers come to the texts with a rough set of topics in mind and want to see which texts fall within these topics, thus shifting the nature of the research more towards confirmatory data analysis. In this case, they can use a "semi-supervised" Labeled LDA algorithm like CorEx (Gallagher *et al.* 2017), which allows them to "suggest" salient topics to the LDA algorithm before it runs. In our French *ancien régime* example, Labeled LDA would allow researchers to suggest single keywords like "impôts" (taxes) or keyword groups like {"gabelle", "taille"} (the salt tax and the land tax, respectively, two of the most onerous *ancien régime* tax

burdens in the eyes of the peasantry), thus nudging the LDA algorithm towards detecting taxation-related topics.

But how exactly are these algorithms able to detect topics within a corpus without being given any domain-specific knowledge whatsoever? This seeming miracle is achieved, in short, via a statistical model of the writing process: since texts tend to be structured—into sentences, paragraphs, and chapters, for example—in such a way that a given section (say, a paragraph) focuses on only one or a few topics selected from the much wider set of topics existing across the corpus, we can infer that the more a pair of words co-occurs *within* these sections the more likely they are to pertain to the same topic. At a high level, therefore, we can scan over a corpus and discover its constituent topics simply by identifying clusters of words that tend to occur close to one another. A topic then, in this sense, is "coherent" to the extent that its words are more likely to co-occur than we would expect by "random chance"—i.e., if the words were randomly selected one-by-one from a dictionary. This captures the intuition that, for example, since the Wikipedia article for Thomas Hobbes is about a particular topic, words which occur *within* this article (choosing eight at random, programmatically) are more likely to be about the same topic than words chosen at random across all words in all Wikipedia articles (again choosing eight at random, programmatically), as illustrated in Table 3.

| **Thomas Hobbes** | **Karl Marx** | **All Wikipedia** |
|---|---|---|
| war | economic | wolf |
| civil | rebellions | chondrite |
| consent | converted | bawdy |
| among | mode | larynx |
| patriarchalists | right-wing | perchance |
| history | countries | cytosine |
| enter | struggle | pollinate |
| government | industrialisation | negligible |

Table 3: Randomly-chosen words from Wikipedia articles on Thomas Hobbes and Karl Marx, versus randomly-chosen words from across all Wikipedia articles

Concretely, for reasons which will be elucidated in the next section, we can transform this intuition into a computationally-estimable model by specifying a "data-generating process"—essentially

a stylized story of how the author "chose" the sequences of words which appear in the texts. For example, previewing what is to come, Algorithm 1 represents the basic data-generating process underlying all topic models.

**Algorithm 1** Data-Generating Process for a Text Corpus

1. *Choose hyperparameters.* The author decides how many documents $N$ they'd like to write, as well as a set of $K$ topics they'd like to write about in these texts. We assume the author has a vocabulary of $M$ words $V = \{w_1, w_2, \ldots, w_M\}$ to draw on while writing.
2. *Specify topics.* For each topic $t \in \{t_1, t_2, \ldots, t_K\}$, the author evaluates which words in $V$ are most and least pertinent. For example, if the topic is astronomy, "planet", "moon", and "orbit" will be given high scores, while "koala" and "headache" will be given low scores. Each topic $t$ is thus assigned a corresponding set of importance scores for each word: $p_t(v)$ represents the importance to topic $t$ of word $v$. In our example, the author may choose $p_t(\texttt{planet}) = p_t(\texttt{moon}) = p_t(\texttt{orbit}) = 0.3$, and $p_t(\texttt{koala}) = p_t(\texttt{headache}) = 0.05$.
3. *Specify documents.* For each document $d \in \{d_1, d_2, \ldots, d_N\}$, the author chooses
   (a) how many words $N_d$ they'd like to write for this document, thus creating blank slots $S_d = \{s_{d,1}, s_{d,2}, \ldots, s_{d,N_d}\}$, and
   (b) the distribution of topics $\theta_d$ they'd like to cover in this document. For example, if $d$ is an article on the Franco-Prussian War, they might choose $\theta_d$ to be a balance between three topics: 40% France, 40% Prussia, and 20% warfare. Mathematically, if the global set of topics $T$ chosen in Step 1 was $t_1$ = France, $t_2$ = Prussia, $t_3$ = warfare, $t_4$ = Buddhism, this balance would be represented by $\theta_d = (0.4, 0.4, 0.2, 0.0)$.
4. *Fill the documents with words.* For each slot $s_{d,i}$ in a document $d$, the author chooses a topic for this slot $t_{d,i}$ based on the document's topic distribution $\theta_d$ (the distribution chosen in the previous step). They then choose a word $w$ from their vocabulary $V$ to fill in this slot, with the probability of choosing a specific word $w_j$ proportional to that word's topic-$t_{d,i}$ importance score $p_{t_{d,i}}(w_j)$.

Despite the simplicity of this model, its definition of a topic as a grouping of words precisely captures the word co-occurrence intuition from above: rather than choosing words at random, as in the All Wikipedia column of Table 3, the author first chooses a *topic*, which thus establishes the likelihood that particular words are chosen. In this way, continuing our example, choosing the Thomas Hobbes topic increases the likelihood that the term "civil" is chosen to be the next word in the text (relative to its frequency in the English language in general), whereas choosing the Karl Marx topic would instead make "industralisation" more likely to be chosen (again, relative to its

overall frequency in English).

Once this data-generating process is specified, all that is left to do for a computer to discover the latent topics within a corpus is to "run it backwards": in essence, to divide the words in the texts up into clusters which best fit this story. Putting "chondrite" and "cytosine" (the first being a type of meteorite, the second a nucleotide found in DNA) into the same cluster, for example, would not fit the story well, since these terms will rarely co-occur in the same section of a text. If it instead placed "chondrite" into a cluster with other astronomy terms and "cytosine" into a separate cluster of microbiological terms, this would better explain the tendency for words in the former to co-occur in texts about astronomy, and for words in the latter to co-occur in texts about microbiology.

The fact that a computer is able to estimate this "best" division in mere seconds (i.e., without trying every single possible split of the vocabulary into separate word clusters, which would be prohibitively time-consuming and computationally expensive) is due to the magic of Probabilistic Graphical Models (PGMs), algorithms which allow researchers to specify the parameters of their data-generating processes—*any* data-generating processes which adhere to a loose set of constraints—and nearly instantly obtain estimates for the parameters within these specifications. In the case of general topic models, for example, the basic parameters would be (a) the distribution of *topics* across the corpus, and (b) the distribution of *words* within each topic. By restricting the total number of topics $K$ that we want the algorithm to generate, we force it to choose a small set of coherent topics, based on the $K$ word clusters with the strongest co-occurrence patterns. In the next section we introduce the specific form that these Probabilistic Graphical Models take, and walk through an example of how they can be constructed around a set of hypotheses.

Every such PGM is simply the formal mathematical representation of a data-generating process, such as the text-generation process described above in Algorithm 1. So, if we wanted to model the relationship between weather and a person's choice of whether to go out and party or stay in and watch a movie on a given Saturday evening, we could begin by proposing the data-generating process given in Algorithm 2.

Now, given the description of a PGM given above (nodes as variables, edges as relationships

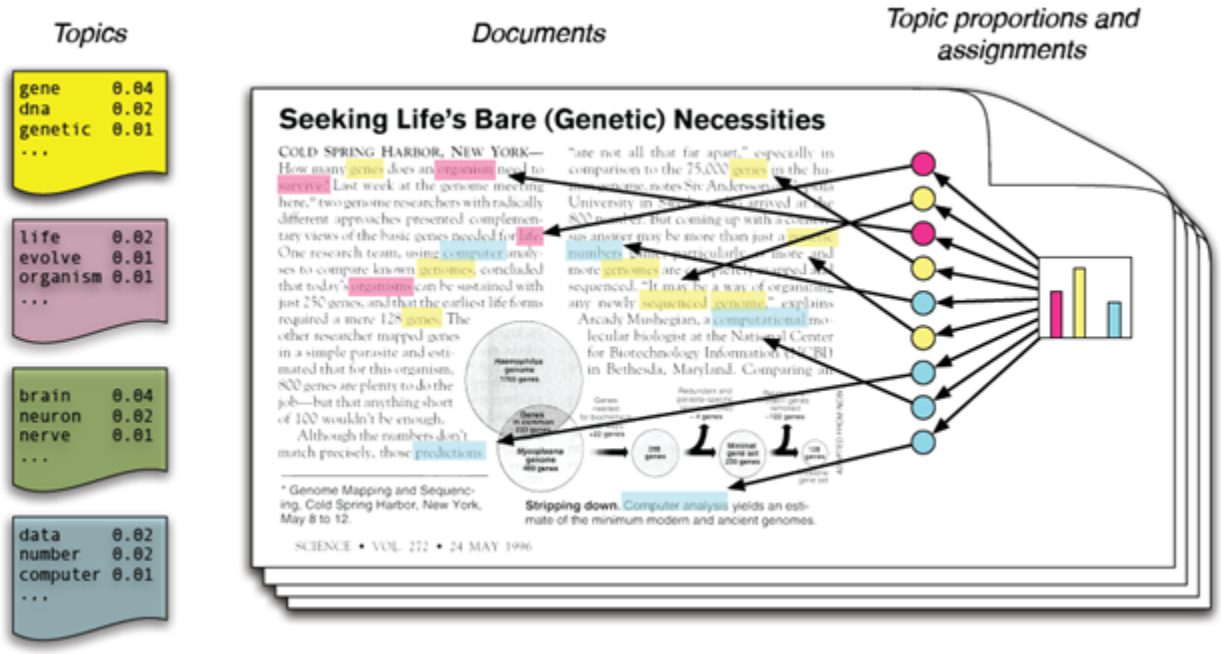



Figure 31: From Blei 2012, p. 3

**Algorithm 2** Data-Generating Process for Evening Plans

1. The person $P$ looks out the window and observes the weather.
2. If the weather is sunny, $P$ goes out to a party. Otherwise, $P$ stays in and watches a movie.

between variables), we can perform the move alluded to in the previous section: we can convert our data-generating process into a PGM, by defining nodes (variables) and edges (relationships) as follows:

1. A variable $w$ which can take on values in $\{\mathsf{Sunny}, \mathsf{Rainy}\}$
2. A variable $a$ which can take on values in $\{\mathsf{Go\ Out}, \mathsf{Stay\ In}\}$, and
3. An edge $e$ from $w$ to $a$ which encodes the intuition that one is more likely to go out if it's sunny than if it's rainy via the probability distribution $P(\mathsf{Go\ Out} \mid \mathsf{Sunny}) = 0.75$, $P(\mathsf{Stay\ In} \mid \mathsf{Sunny}) = 0.25$, $P(\mathsf{Go\ Out} \mid \mathsf{Rainy}) = 0.25$, and $P(\mathsf{Stay\ In} \mid \mathsf{Rainy}) = 0.75$.

The resulting PGM, in graphical form[67], is presented in Figure 32, where the Conditional

[67]The "Graphical" in Probabilistic Graphical Model is not used in the same sense as the "graphical" we're used

Probability Table describing the edge from the $w$ node to the $a$ node is given in Table 4.

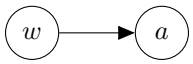



Figure 32: A basic PGM, representing the relationship between $w$, the weather, and $a$, the subsequent action of a person deciding whether to go out or stay in for the night.

| | Go Out | Stay In |
|---|---|---|
| Sunny | 0.75 | 0.25 |
| Rainy | 0.25 | 0.75 |

Table 4: The Conditional Probability Table for the PGM shown in Figure 32.

PGMs can help us make inferences about the world in the face of incomplete information, which is the situation in nearly every real-world problem. The key tool here is the separation of nodes into two categories: *observed* (represented graphically as a shaded node) and *latent* (represented graphically as an unshaded node). Thus we can now use our model as a weather-inference machine: if we observe that the person we're modeling is out at a party with us, what can we infer from this information about the weather outside? We can draw this situation as a PGM with shaded and unshaded nodes, as in Figure 33, and then use Bayes' Rule to perform calculations over the network, to see how the observed information about the person at the party "flows" back into the node representing the weather.

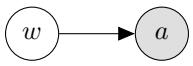



Figure 33: A PGM representing the same situation as in Figure 32, except that the node for variable $a$ is now shaded, indicating a situation where we have observed the person's action ($a = \text{Go Out}$) but still only have a probability distribution over the weather $w$.

Keeping in mind that Bayes' Rule tells us, for any two events $A$ and $B$, how to use information about $P(B|A)$ to obtain information about $P(A|B)$:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

to from vernacular English. Capital-G Graphical denotes that the Probabilistic Model is represented as a Graph, a well-defined mathematical object consisting of nodes and edges, which does not have to be represented graphically (though it could be, like in our example here with circles and arrows). In fact, when a computer program is estimating a PGM, it is by definition not in a graphical form—it's in the form of 0s and 1s, stored in the computer's memory.

We can now apply this rule to obtain our new probability distribution over the weather, taking into account the new information that the person has chosen to go out:

$$\begin{aligned} P(w = \textsf{Sunny} \mid a = \textsf{Go Out}) &= \frac{P(a = \textsf{Go Out} \mid w = \textsf{Sunny})}{P(a = \textsf{Go Out})} \\ &= \frac{P(a = \textsf{Go Out} \mid w = \textsf{Sunny})}{P(a = \textsf{Go Out} \mid w = \textsf{Sunny}) + P(a = \textsf{Go Out} \mid w = \textsf{Rainy})} \end{aligned}$$

And now we simply plug in the information we already have from our conditional probability table to obtain our new (conditional) probability of interest:

$$P(w = \textsf{Sunny} \mid a = \textsf{Go Out}) = \frac{(0.8)(0.5)}{(0.8)(0.5) + (0.1)(0.5)} = \frac{0.4}{0.4 + 0.05} = \frac{0.4}{0.45} \approx 0.89.$$

We have learned something interesting: now that we've observed the person out at a party, the probability that it is sunny out jumps from 0.5 (called the "prior" estimate of $w$, i.e., our best guess without any other relevant information) to 0.89 (called the "posterior" estimate of $w$, i.e., our best guess after incorporating relevant information).

### A.3 Topic Models

Turning specifically to the analysis of historical *texts* using PGMs, we can now understand probabilistic *topic models* in a similar fashion to the weather-estimation model: note that, in the previous example, we took information about an *observed* quantity—the presence or absence of someone at our party—and used it to draw inferences about an *unobserved* quantity—the weather. Analogously, for our topic model, we consider the words in the text to be our *observed* data, and try to use this observed data to draw inferences about the *unobserved* topics underlying the choice of words in a given text. Just as we used the presence of someone at the party to infer a higher likelihood that it's sunny out, our topic models will use the presence of particular words in a *cahier*—say, "droits", "liberté", "égalité"—to infer a higher likelihood that this grievance pertains to a particular topic—in this case, a topic involving the concept of "individual rights".

Given these parallels between our previous example and the basic structure of a topic model, then, we can start to model the data-generating process for a single word in the text, as illustrated

in Figure 34. Note that it is identical to our example PGM in Figure 33, except for the labeling of the variables: the node which represented the observed action $a$ now represents the observed word $w_{d,i}$ (the $i$th word in document $d$), and the node which represented the unobserved weather $w$ now represents the unobserved topic $t_{d,i}$.

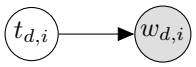



Figure 34: A first attempt at a PGM representing the data-generating process for an observed word $w_{d,i}$, the $i$th word in document $d$, within a text.

However, by comparing this PGM with our full data-generating process in Algorithm 1 above, we can see that the former is not sufficient for representing the mechanics of this process. Whereas in the evening-plans model we *assumed* specific values for the conditional probabilities of each action—for example, that $P(\textsf{Go Out} \mid \textsf{Sunny}) = 0.75$—in the topic modeling case we want to *estimate* (rather than assume) the probabilities of words within each topic itself. In other words, we now have two "layers" of unknown quantities: we want to estimate *both* the document-topic distributions $\theta_d$ *and* the topic-word distributions $p_t(w)$.

The power of modeling via PGMs becomes apparent here since, just as we were able to split the full data-generating process in Algorithm 1 into steps, we can start building towards the full PGM by constructing "sub-PGMs", each corresponding to one step in the data-generating process. To start with, we can consider what the PGM for the Specify-Documents sub-process would look like, and arrive at a sub-PGM like the one illustrated in Figure 35.

The node for the observed word $w$ is in dotted rather than solid outline here to denote the fact that this PGM is *incomplete* on its own: it represents the step in which the author specifies the relationships represented by edges here, i.e., specifies the likelihood of choosing a given word for each of the $K$ topics. An example of such a word-topic specification, which would complete the definition of this sub-PGM, is given in Table 6. But the dotted-outline $w$ node here is *not* the same as the eventual observed-word node of our full data-generating process, since in this full process the generation of a word depends on *both* the likelihood of a word for each topic (as modeled here) *and* the particular topic which was chosen for the slot in which the word will appear. In other words, it is only when the author has access to both of these pieces of information—when they have (a)

the Conditional Probability Table specifying how to choose a topic for a given slot, and (b) the Conditional Probability Table specifying how to choose a word from that chosen topic—that they can proceed to generating the word itself. To address this incompleteness, we can *separately* model the process by which the topic for each word-slot is chosen, the Specify-Documents sub-process of our full data-generating process, as illustrated in Figure 35.

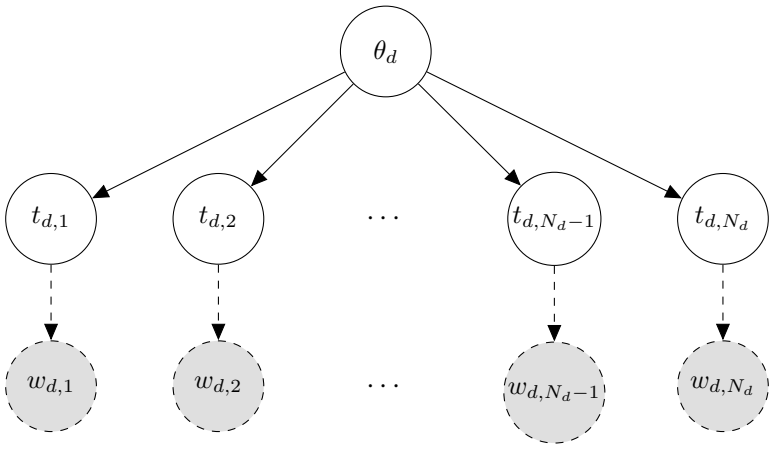



Figure 35: A PGM representing the Specify-Documents sub-process, in which the author specifies a topic distribution $\theta_d$.

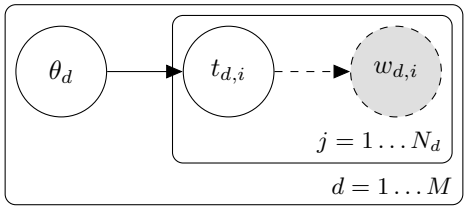



Figure 36: A PGM representing the Specify-Documents sub-process, using plate notation to condense the repetition in Figure 35.

| | **Document** | | | |
|---|---|---|---|---|
| **Topic** | Document 1 | Document 2 | $\cdots$ | Document $M$ |
| Astronomy | 0.80 | 0.25 | $\cdots$ | 0.10 |
| Marsupials | 0.20 | 0.25 | $\cdots$ | 0.00 |
| Music | 0.00 | 0.50 | $\cdots$ | 0.00 |
| Vegetables | 0.00 | 0.00 | $\cdots$ | 0.90 |

Table 5: An example Conditional Probability Table for the Specify-Documents PGM shown in Figures 35 and 36.

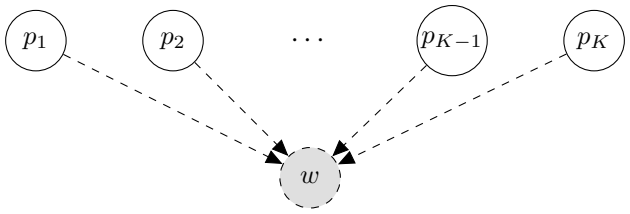



Figure 37: A PGM representing the Specify-Topics sub-process, in which an author specifies a word importance score $p_t(w)$ for each word $w$ and topic $t$.

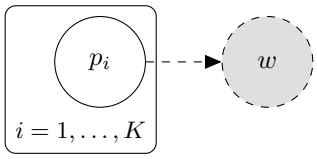



Figure 38: A PGM representing the Specify-Topics sub-process, using plate notation to condense the repetition in Figure 37.

| | **Topic** | | | |
|---|---|---|---|---|
| **Word** | Astronomy | Marsupials | Music | Vegetables |
| planet | 0.30 | 0.03 | 0.01 | 0.06 |
| moon | 0.30 | 0.03 | 0.01 | 0.06 |
| orbit | 0.30 | 0.03 | 0.01 | 0.06 |
| koala | 0.02 | 0.41 | 0.01 | 0.06 |
| possum | 0.02 | 0.41 | 0.01 | 0.06 |
| guitar | 0.02 | 0.03 | 0.93 | 0.06 |
| spinach | 0.02 | 0.03 | 0.01 | 0.32 |
| broccoli | 0.02 | 0.03 | 0.01 | 0.32 |

Table 6: An example Conditional Probability Table for the Specify-Topics PGM shown in Figure 37.

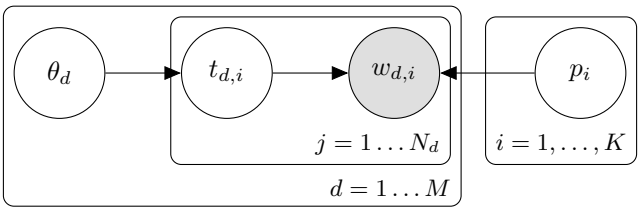



Figure 39: The complete PGM representing the data-generating process for an observed word $w_{d,i}$, with the Specify-Documents and Specify-Topics sub-processes incorporated explicitly.

In words, this PGM diagram tells us that given a choice of topic $t_{d,i}$ chosen for slot $s_{d,i}$, *in conjunction with* a choice of word-importance scores $p_t(w)$ for all words $w$ relative to topic $t$, we can determine exactly the probability of the word $w_{d,i}$ appearing in slot $i$ of document $d$.

This walkthrough of constructing a PGM from scratch has illustrated two powerful properties of PGMs that aided our modeling process: on the one hand, we have seen how *factoring* the full data-generating process into sub-processes allowed us to ignore the details of e.g. the Specify-Topics subprocess while we thought through the Specify-Documents subprocess, and then gave us a natural way to join these parts together into a fully-specified model of the whole (by joining the two PGMs at the $w$ node, which both sub-processes had only partially specified). Returning to our discussion of complex historical and social processes in the beginning of the chapter, one can see how effective this factoring-and-rejoining method could be for rendering studies of complex phenomena more manageable: a team of researchers could establish a division of labor, with each individual researcher modeling their particular sub-phenomena of interest or expertise via PGMs, then join their respective PGMs at particular nodes of intersection. For example, in the French Revolution case, a node representing the text of debates in the National Assembly could sit at the intersection not only of individual models of the three estates, but also a model of the Parisian crowd who often swayed these debates via cheering or booing from the gallery (a dynamic which George Rudé's *The Crowd in the French Revolution* (Rudé 1959) brought to the forefront of French Revolution research[68]), as illustrated in Figure 40.

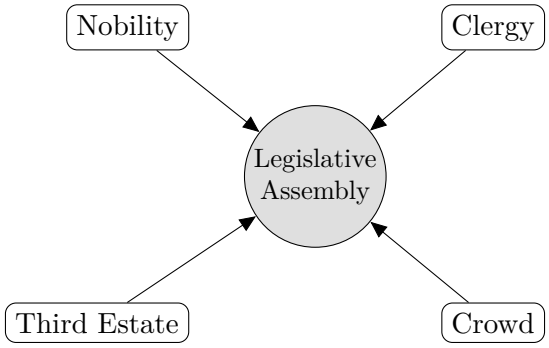



Figure 40: A PGM representing the fusion of several individual models of French Revolutionary entities, to explain the observed outcomes in the revolutionary Legislative Assembly.

[68]See also Rudé's subsequent work, *The Crowd in History* (Rudé 1964), for an even broader examination of the role of the crowd in explaining historical social behavior.

## References


[1] T. Adorno, E. Frenkel-Brenswik, D. J. Levinson, and R. N. Sanford, *The Authoritarian Personality*. Verso Books, 1950.

[2] L. Althusser, *For Marx*. Verso, 1968.

[3] H. Arendt, *The Origins of Totalitarianism*. Schocken Books, 1951.

[4] E. Ash, D. L. Chen, and S. Naidu, "Ideas Have Consequences: The Effect of Law and Economics on American Justice," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 2992782, 2017.

[5] E. Ash, J. Jacobs, B. MacLeod, S. Naidu, and D. Stammbach, "Unsupervised Extraction of Workplace Rights and Duties from Collective Bargaining Agreements," *2020 International Conference on Data Mining Workshops (ICDMW)*, 2020, 766–774.

[6] R. Ashcraft, *Locke's Two Treatises of Government*. Routledge, 1986.

[7] J. L. Austin, *How to Do Things with Words*. Clarendon Press, 1962.

[8] S. Avineri, *The Social and Political Thought of Karl Marx*. Cambridge: Cambridge University Press, 1968.

[9] A. T. J. Barron, J. Huang, R. L. Spang, and S. DeDeo, "Individuals, Institutions, and Innovation in the Debates of the French Revolution," *Proceedings of the National Academy of Sciences*, vol. 115, no. 18, 4607–4612, 2018.

[10] T. Bayes, "An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton," *Philosophical transactions of the Royal Society of London*, no. 53, 370–418, 1763.

[11] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," *arXiv:2004.05150*, 2020.

[12] M. Bevir, "Are there Perennial Problems in Political Theory?" *Political Studies*, vol. 42, no. 4, 662–675, 1994.

[13] M. Bevir, *The Logic of the History of Ideas*. New York: Cambridge University Press, 1999.

[14] D. M. Blei, "Introduction to Probabilistic Topic Models," *Communications of the ACM*, vol. 55, no. 4, 77–84, 2012.

[15] E. A. Chi, J. Hewitt, and C. D. Manning, "Finding Universal Grammatical Relations in Multilingual BERT," *arXiv:2005.04511 [cs]*, 2020.

[16] G. A. Cohen, *Karl Marx's Theory of History: A Defence*. Princeton University Press, 1978.

[17] R. G. Collingwood, *The Idea of History*. Oxford: Clarendon Press, 1946.

[18] B. Comrie, *The Languages of the Soviet Union*. CUP Archive, 1981.

[19] W. E. Connolly, *The Terms of Political Discourse*. Princeton University Press, 1974.

[20] R. Darnton, D. Roche, and N. Y. P. Library, *Revolution in Print: The Press in France, 1775-1800*. University of California Press, 1989.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv:1810.04805 [cs]*, 2019.

[22] I. Dokmanic, R. Parhizkar, J. Ranieri, and M. Vetterli, "Euclidean Distance Matrices: Essential theory, algorithms, and applications," *IEEE Signal Processing Magazine*, vol. 32, no. 6, 12–30, 2015.

[23] M. Dummett, *Frege: Philosophy of Language*. Harvard University Press, 1973.

[24] E. Dussel, *Towards An Unknown Marx: A Commentary on the Manuscripts of 1861-63*. Routledge, 2002.

[25] J. Elster, "The Case for Methodological Individualism," *Theory and Society*, vol. 11, no. 4, 453–482, 1982.

[26] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT Sentence Embedding," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland: Association for Computational Linguistics, 2022, 878–891.

[27] L. S. Feuer, "The North American Origin of Marx's Socialism," *Western Political Quarterly*, vol. 16, no. 1, 53–67, 1963.

[28] J. R. Firth, *Papers in Linguistics, 1934-1951*. Oxford University Press, 1957.

[29] R. J. Gallagher, K. Reing, D. Kale, and G. V. Steeg, "Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge," *Transactions of the Association for Computational Linguistics*, vol. 5, no. 0, 529–542, 2017.

[30] D. R. Gandy, *Marx and History: From Primitive Society to the Communist Future*. University of Texas Press, 1979.

[31] A. Gerow, Y. Hu, J. Boyd-Graber, D. M. Blei, and J. A. Evans, "Measuring Discursive Influence Across Scholarship," *Proceedings of the National Academy of Sciences*, vol. 115, no. 13, 3308–3313, 2018.

[32] A. Giddens, *Central Problems in Social Theory: Action, Structure, and Contradiction in Social Analysis*. University of California Press, 1979.

[33] D. Goodman, *The Republic of Letters: A Cultural History of the French Enlightenment*. Cornell University Press, 1996.

[34] A. Gourevitch, *From Slavery to the Cooperative Commonwealth: Labor and Republican Liberty in the Nineteenth Century*. Cambridge: Cambridge University Press, 2015.

[35] H. Grégoire, R. Jakobson, and M. Szeftel, *La geste du prince Igor, épopée russe du douzième siècle*. New York: Institut de philologie et d'histoire orientales et slaves, 1948.

[36] J. C. Harsanyi, "Cardinal Utility in Welfare Economics and in the Theory of Risk-taking," *Journal of Political Economy*, vol. 61, no. 5, 434–435, 1953.

[37] K. Heffernan, O. Çelebi, and H. Schwenk, *Bitext Mining Using Distilled Sentence Representations for Low-Resource Languages*, 2022.

[38] K. Hlaváčková-Schindler, *The Assumption of Non-Gaussianity in Natural and Social Sciences and Its Influence on Detection of Causal Relationships*. IntechOpen, 2012.

[39] K. J. Hudson, "The Double Blow: 1956 and the Communist Party of Great Britain," Doctoral, University of London, 1992.

[40] J. Jacobs, "Quantifying Cultural Diplomacy: The Translation and Diffusion of Marxism from the Communist Manifesto to the Cold War," *Working Paper*, 2021.

[41] R. Jakobson, *Child Language, Aphasia and Phonological Universals*. Walter de Gruyter, 1941.

[42] R. Jakobson, "Linguistics in Its Relations to Other Sciences," *Eight Decades of General Linguistics*, 265–304, 1957.

[43] R. Jakobson, "Linguistics and Poetics," *Linguistics and Poetics*, De Gruyter Mouton, 1960, 18–51.

[44] D. Jurafsky, *The Language of Food: A Linguist Reads the Menu*. W. W. Norton & Company, 2014.

[45] A. C. Kozlowski, M. Taddy, and J. A. Evans, "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings," *American Sociological Review*, vol. 84, no. 5, 905–949, 2019.

[46] V. Kulkarni, R. Al-Rfou, B. Perozzi, and S. Skiena, "Statistically Significant Detection of Linguistic Change," *Proceedings of the 24th International Conference on World Wide Web*, Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2015, 625–635.

[47] A. Lazaridou, E. Bruni, and M. Baroni, "Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world," *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, Maryland: Association for Computational Linguistics, 2014, 1403–1414.

[48] N. Levine, *Divergent Paths: The Hegelian Foundations of Marx's Method*. Lanham, MD: Lexington Books, 2006.

[49] M. Lewis-Beck, A. E. Bryman, and T. F. Liao, *The SAGE Encyclopedia of Social Science Research Methods*. SAGE Publications, 2003.

[50] J. A. London, "Re-imagining the Cambridge School in the Age of Digital Humanities," *Annual Review of Political Science*, vol. 19, no. 1, 351–373, 2016.

[51] Y. Ma, S. Mukherjee, and B. Uzzi, "Mentorship and protégé success in STEM fields," *Proceedings of the National Academy of Sciences*, vol. 117, no. 25, 14077–14083, 2020.

[52] K. Marx, *Herr Vogt*. A. Petsch & Company, 1860.

[53] R. McElreath, *Statistical Rethinking: A Bayesian Course with Examples in R and STAN*. CRC Press, 2020.

[54] D. McLellan, *Marxism After Marx*. New York: Palgrave Macmillan, 2007.

[55] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," *arXiv:1310.4546 [cs, stat]*, 2013.

[56] F. Morton, *Thunder at Twilight: Vienna 1913/1914*. Collier Books, 1990.

[57] F. Mosteller and D. L. Wallace, *Inference and Disputed Authorship: The Federalist*. Center for the Study of Language and Information, 1964.

[58] B. Parekh and R. N. Berki, "The History of Political Ideas: A Critique of Q. Skinner's Methodology," *Journal of the History of Ideas*, vol. 34, no. 2, 163–184, 1973.

[59] J. Plamenatz, *Man and Society: Political and Social Theories from Machiavelli to Marx*. Longman, 1963.

[60] J. G. A. Pocock, *Virtue, Commerce, and History: Essays on Political Thought and History, Chiefly in the Eighteenth Century*. Cambridge: Cambridge University Press, 1985.

[61] J. G. A. Pocock, *Political Thought and History: Essays on Theory and Method*. Cambridge University Press, 2009.

[62] J. G. A. Pocock, *The Machiavellian Moment: Florentine Political Thought and the Atlantic Republican Tradition*. Princeton University Press, 1975.

[63] K. Popper, *The Open Society and Its Enemies*. Routledge, 1945.

[64] S. S. Prawer, *Karl Marx and World Literature*. Oxford: Clarendon Press, 1976.

[65] W. V. O. Quine, *Word and Object*. Cambridge, MA: MIT Press, 1960.

[66] J. Rawls, "Outline of a Decision Procedure for Ethics," *The Philosophical Review*, vol. 60, no. 2, 177–197, 1951.

[67] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *arXiv:1908.10084*, 2019.

[68] L. Rheault and C. Cochrane, "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora," *Political Analysis*, vol. 28, no. 1, 112–133, 2020.

[69] J. E. Roemer, *Theories of Distributive Justice*. Harvard University Press, 1996.

[70] M. A. Rose, *Reading the Young Marx and Engels: Poetry, Parody, and the Censor*. Croom Helm, 1978.

[71] G. Rudé, *The Crowd in the French Revolution*. Oxford U.P. Paperback, 1959.

[72] G. Rudé, *The Crowd in History: A Study of Popular Disturbances in France and England, 1730-1848*. Serif, 1964.

[73] F. de Saussure, *Course in General Linguistics*. Open Court, 1916.

[74] A. M. Schlesinger, *The Vital Center: The Politics of Freedom*. Houghton Mifflin Company, 1949.

[75] J. W. Scott and D. Keates, *Schools of Thought: Twenty-five Years of Interpretive Social Science*. Princeton University Press, 2001.

[76] G. Shapiro, J. Markoff, and S. R. Duncan Baretta, "The Selective Transmission of Historical Documents: The Case of the Parish Cahiers of 1789," *Histoire & Mesure*, vol. 2, no. 3, 115–172, 1987.

[77] Q. Skinner, "Meaning and Understanding in the History of Ideas," *History and Theory*, vol. 8, no. 1, 3–53, 1969.

[78] Q. Skinner, *The Foundations of Modern Political Thought: Volume 1, The Renaissance*. Cambridge University Press, 1978.

[79] Q. Skinner, *The Foundations of Modern Political Thought: Volume 2, The Age of Reformation*. Cambridge University Press, 1978.

[80] Q. Skinner, *The Return of Grand Theory in the Human Sciences*. Cambridge University Press, 1990.

[81] Q. Skinner, *Liberty Before Liberalism*. Cambridge University Press, 1998.

[82] Q. Skinner, *Hobbes and Republican Liberty*. Cambridge University Press, 2008.

[83] Q. Skinner, *Visions of Politics: Volume 1, Regarding Method*. Cambridge University Press, 2012.

[84] A. Sokal and J. Bricmont, *Fashionable Nonsense: Postmodern Intellectuals' Abuse of Science*. Picador, 1997.

[85] S. Soni, L. F. Klein, and J. Eisenstein, "Abolitionist Networks: Modeling Language Change in Nineteenth-Century Activist Newspapers," *Journal of Cultural Analytics*, vol. 6, no. 1, 18841, 2021.

[86] D. Sperber, *Explaining Culture: A Naturalistic Approach*. Cambridge: Blackwell, 1996.

[87] L. Strauss, *What Is Political Philosophy? And Other Studies*. Chicago: University of Chicago Press, 1959.

[88] J. L. Talmon, *The Rise of Totalitarian Democracy*. Beacon Press, 1952.

[89] C. Tilly, *Contentious Performances*. New York: Cambridge University Press, 2008.

[90] J. E. Toews, *Hegelianism: The Path Toward Dialectical Humanism, 1805-1841*. Cambridge University Press, 1985.

[91] J. Tully, *Meaning and Context: Quentin Skinner and His Critics*. Princeton University Press, 1988.

[92] M. Vendruscolo, E. Kussell, and E. Domany, "Recovery of protein structure from contact maps," *Folding and Design*, vol. 2, no. 5, 295–306, 1997.

[93] C. Welch, J. K. Kummerfeld, V. Pérez-Rosas, and R. Mihalcea, "Exploring the Value of Personalized Word Embeddings," *Proceedings of the 28th International Conference on Computational Linguistics*, International Committee on Computational Linguistics, 2020, 6856–6862.

[94] M. Wevers, J. Gao, and K. L. Nielbo, "Tracking the Consumption Junction: Temporal Dependencies between Articles and Advertisements in Dutch Newspapers," *Digital Humanities Quarterly*, vol. 014, no. 2, 2020.

[95] M. Wevers and M. Koolen, "Digital begriffsgeschichte: Tracing semantic change using word embeddings," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, vol. 53, no. 4, 226–243, 2020.

[96] J. Williamson, *In Defence of Objective Bayesianism*. Oxford: Oxford University Press, 2010.

[97] L. Wittgenstein, *Tractatus Logico-Philosophicus: Centenary Edition*. Anthem Press, 1921.

[98] L. Wittgenstein, *Philosophical Investigations*. Wiley, 1953.

[99] V. Wohl, "Plato Avant la Lettre: Authenticity in Plato's Epistles," *Ramus*, vol. 27, no. 1, 60–93, 1998.

[100] E. M. Wood, *Citizens to Lords: A Social History of Western Political Thought from Antiquity to the Late Middle Ages*. Verso Books, 2008.

[101] J. K. Wright, *Schoenberg, Wittgenstein and the Vienna Circle*. Peter Lang, 2006.

[102] Y. Yang and F. Feng, "Language-Agnostic BERT Sentence Embedding," *Google AI Blog*, 2020.

[103] N. R. Yetman, "The background of the slave narrative collection," *American Quarterly*, vol. 19, no. 3, 534–553, 1967.