MODELING THE LEADERSHIP OF LANGUAGE CHANGE FROM DIACHRONIC TEXT

A Dissertation
Presented to
The Academic Faculty

By

Sandeep Soni

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Interactive Computing

Georgia Institute of Technology

August 2021

MODELING THE LEADERSHIP OF LANGUAGE CHANGE FROM DIACHRONIC TEXT

Approved by:

Dr. Jacob Eisenstein, Advisor
School of Interactive Computing
Georgia Institute of Technology

Dr. Munmun De Choudhury
School of Interactive Computing
Georgia Institute of Technology

Dr. Chao Zhang
School of Computational Science and Engineering
Georgia Institute of Technology

Dr. Rada Mihalcea
Department of Electrical Engineering
University of Michigan

Dr. David Bamman
School of Information
University of California, Berkeley

Date Approved: June 25, 2021

Time changes all things; there is no reason why language should escape this universal law

Ferdinand de Saussure

In the memory of my dad

ACKNOWLEDGEMENTS

When I started working towards this thesis, I could not believe my luck of being in a fully funded PhD program at a premier US university with the opportunity to do interdisciplinary research. There are so many to thank for making this thesis possible in the first place and then helping it take shape.

A huge behind-the-scenes effort is required to help any student focus on research — my case is no different. I want to start by acknowledging the work of Jessica Celestine, Renee Jamieson, Tina Charest, Danielle Shenise, and Becky Wilson. They were responsible for handling several queries ranging from admissions to visa approvals, from travel reimbursements to getting teaching assistantships, and many others that I don't remember now. The school of interactive computing has been a smooth working machine because of the labor of these women which should not go unnoticed. A special thanks also to Gregory Abowd who was the PhD coordinator when I started in the program. Gregory conducted a series of seminars that were highly informative and reassuring as I took baby steps in PhD research. I'm also very grateful to Rosa Arriaga, who later became the PhD coordinator, for her earnest efforts to check my progress without putting too much pressure; the fact that she did this for all the students in the chaos caused due to the pandemic is really remarkable.

Equally important throughout the thesis has been the constant exchange of ideas made possible due to the environment at Georgia Tech. I thank everyone who participated in the social computing seminar hosted by Amy Bruckman, Munmun De Choudhury, and Eric Gilbert. If anyone who attended the seminar with me reads this, I'm sorry to be *that* person who insisted on getting the discussion back on track. Along the same line, I was also fortunate to attend meetings discussing research with the CLAWS group, overseen by Srijan Kumar, where at times I was able to present my own research. I thank everyone from the CLAWS group for giving me feedback on my research ideas.

On the same note, I'm so grateful to members of the Computational Linguistics Lab where I spent all my tenure, first as a masters student and then as a PhD student. I want to thank Rahul Goel for not just being an excellent collaborator but also a good badminton partner and a friend; Yangfeng Ji, Yi Yang, and Umashanthi Pavalanathan for setting the standard for research and inculcating good research habits for the rest of us to follow; Yuval Pinter and Sarah Wiegreffe for bringing fresh ideas to research discussions and being excellent company. A special mention and thanks to Ian Stewart who was my constant friend and companion throughout the PhD. Not only did Ian and I brainstorm technical ideas (helped by the fact that we both worked on similar topics), we connected with our shared love of pop culture. I rolled my eyes every time he improvised a pun, but that was nothing compared to my years of whining about research and life that he had to endure.

I have no doubt that this thesis would not have been possible without the mentoring and support of my advisor, Jacob Eisenstein. Not only did Jacob introduce me to the topic of this thesis, he gently guided me through the highs and lows that came along while I undertook every study. He has helped immensely with the technical ideas in this thesis and been supportive and empathetic with everything in my personal life. I learned from him both in how to think about research but also many life lessons along the way. It has been a real honor for me to be his student and I am fully indebted to him for being my advisor. I simply cannot thank him enough!

This thesis is also a product of multiple collaborations over the years. For that I want to thank Naman Goyal, John Paparrizos, Fernando Diaz, and Hanna Wallach who helped with the social media data acquisition, curation, modeling, and analysis. I thank Shawn Ling Ramirez to help us understand and interpret the cosponsorship data for one of the projects. I thank Kristina Lerman for offering me an internship to work on semantic change, which ultimately proved to be a perfect pivot for this thesis. And Lauren Klein whose long collaboration has been both fruitful, productive, and came at a time when I was at the crossroads with this thesis. If I could be half as good a writer as Lauren then I would count

myself lucky.

I also want to sincerely thank my friends for their encouragement and support: Aishwarya Natarajan, for checking in with me during tough times and being up for a walk whenever I wanted to take a break; Mukul Sati, for being a wonderful roommate and always ready to have a technical brainstorm session; Nishant Yedla, also for being an excellent roommate and cooking for me when I did not have the time; Vivek Sharma and Mahesh Bingi for the gentle words of encouragement and offering a non-research perspective whenever I needed.

Finally, I want to thank my family who made many sacrifices while I took on this intellectual pursuit. My mom, uncle and aunt, brother, sister-in-law, and my late father, shielded me from any distractions. I'm so grateful for their generosity and kindness over the years. Without that support, it's hard to imagine that I could have pulled this thesis off.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Natural languages constantly change over time. These changes are modulated by social factors such as influence which are not always directly observable. However, large-scale computational modeling of language change using timestamped text can uncover the latent organization and social structure. In turn, the social dynamics of language change can potentially illuminate our understanding of innovation, influence, and identity: Who leads? Who follows? Who diverges?

This thesis contributes to the growing body of research on using computational methods to model language change with a focus on quantifying linguistic leadership of change. A series of studies highlight the unique contributions of this thesis: methods that scale to huge volumes of data; measures that quantify leadership at the level of individuals or in aggregate; and analysis that links linguistic leadership to other forms of influence. First, temporal and predictive models of event cascades on a network of millions of Twitter users are used to show that lexical change spreads in the form of a contagion and influence from densely embedded ties is crucial for the adoption of non-standard terms. A Granger-causal test for detecting social influence in event cascades on a network is then presented, which is robust to both the presence of confounds such as homophily and can be applied to model both linguistic or non-linguistic change in a network. Next, a novel scheme to score and identify documents that lead semantic change in progress is introduced. This linguistic measure of influence on the documents is strongly predictive of their influence in terms of the number of citations that they receive for both US court opinions and scientific articles. Subsequently, a measure of lead on any semantic change between a pair of document sources (e.g. newspapers) and a method to aggregate multiple lead-lag relationships into a network is presented. Analysis on an induced network of nineteenth century abolitionist newspapers, following the proposed method, reveals the important yet understated role of women and Black editors in shaping the discourse on abolitionism. Finally, a method to

induce an aggregate semantic leadership network using contextual word representations is proposed to investigate the link between semantic leadership and influence in the form of citations among publication venues that are part of the Association of Computational Linguistics. Taken together, these studies illustrate the utility of finding leaders of language change to gain insights in sociolinguistics and for applications in social science and digital humanities.

CHAPTER 1

INTRODUCTION

An inescapable fact about all active natural languages is that they change (Labov 2001). From sounds (Zeller 1997) to morphology (Anderson 2015), from syntax (Roberts 2007) to semantics (Traugott et al. 2001), languages change at multiple levels. Words, often the most recognizable linguistic units in natural languages, are not immune to changes. New words get added to the inventory of a language (Brinton et al. 2005); for example, the verb *googling* has only become a part of our lexicon after the internet search engine made it ubiquitous. Existing words get replaced by new words; for example, the word *radiogram* was once used synonymously with *X-ray* but has now mostly fallen out of favor (Petersen et al. 2012). Words go out of existence or become obsolete; the term *nuncheon*, for example, referred to a mid-morning snack but is now scarcely used (Keller 1994). Words also shift in meaning; the history of the word *girl* shows it was originally ungendered but only became exclusively referred to a young female since the 15th century (Bybee 2015; Dubossarsky 2018). Countless such examples and overwhelming evidence from research makes one thing abundantly clear: the universality of language change is indisputable.

Discovering *what* has changed in a language is a vast research enterprise, one which not only builds our understanding of the progression of language but also of the societal and cultural evolution from a linguistic perspective (Lehmann 2013). Alternatively, and perhaps more importantly, finding *who* is leading or lagging language change brings to fore the underlying social factors that mediate the process of change. The manifest changes in language are a product of multiple latent factors such as identity, demographic attributes, level of contact, geography, and social networks (e.g., Labov 2001; Trudgill 2000; Wilson et al. 2011). Focusing solely on the social question of language change offers a lens to deepen our knowledge about the social structure and organization, which may also not be directly observable; it can help in identifying innovators, adopters, influencers, and resistors of language change. This thesis investigates the social dynamics of language change as it pertains to linguistic leadership, with a particular emphasis on combining computational methods from network and text analysis to quantify leadership.

A core, well-established concern of sociolinguistics — the research area within linguistics that studies the interplay between language and society — is the ability to find leaders of language change. This thesis takes inspiration from the traditional sociolinguistic research paradigm but differs in four key respects. While sociolinguists have outlined the determinants of change, much of their research is concentrated in investigating sound or morphological change. In contrast, this thesis studies changes in frequency and meaning of words. The linguistic variables in sociolinguistic studies are typically limited, high frequency or *a prioiri* known. This thesis, on the other hand, applies methods on large amounts of text that can unearth changes that are rare and previously unknown. Under the traditional paradigm, most studies analyze changes in *apparent time* — that is, by comparing the language patterns of individuals from different age groups at the same time. Instead, this thesis analyzes timestamped text to detect changes in *real time* — that is, by comparing the usage at two different times (Bailey 2002). Finally, with respect to the study of language change in a social network, sociolinguistic studies have considered networks of small, local networks obtained by snowball sampling. In contrast, this thesis uses network methods to model change on networks of size as small as a few dozen to as large as a few million nodes, without restricting to a local ego-centric view of a network that may lead to biases inferences.

The quest of finding leaders of language change is not restricted to sociolinguistics alone. Language change can serve as a proxy to other changes; for example, the increase in frequency of terms such as *web* and *cell* in the last three decades is correlated with changes and the widespread adoption of communication and information technology. Finding the influential actors behind language change can equate to finding innovators whose ideas were picked up or ignored (e.g. Barron et al. 2018). Language change can thus be an instrument for finding and evaluating the impact of cultural innovations and the leaders of cultural changes.

Finding linguistic leaders of language change is not short of methodological challenges. One should ideally have methods that can automatically make inferences about linguistic leadership without running into the limitations of scale. Methods should also be equipped at finding leaders at different granularity: at the level of an individual change, a set of changes, among individuals, or aggregating individuals into groups. This methodological flexibility can help in assessing the influence of linguistic leaders more holistically: from how language spreads through interpersonal influence along network ties for individual changes to the global role that individual leaders play in language change when aggregated over multiple changes.

Once again, the methodological challenges are extended to disciplines other than sociolinguistics. For example, in digital humanities, distant reading is a mode of macro-analysis on a large amount of data using computational methods that often requires making aggregate inferences (Moretti 2000; Underwood 2017). As a further point, scalable methods to find leaders or influencers in, for example, the transmission of lexical innovations (as described in the studies of chapter 3 and chapter 4) can be applied to the spread of other types of innovations such as memes (Leskovec et al. 2009; Gomez-Rodriguez et al. 2012).

The desiderata then is the following: find leaders of language change from timestamped text by devising methods that scale to large amounts of data; have the ability to aggregate leadership over multiple changes; and make inferences about linguistic leadership at individual or group level. No single method is fully equipped to satisfy all the requirements. This thesis proposes a series of computational models that can all operate on diachronic text to satisfy the broad aforementioned desiderata.

The rest of the chapter is structured as follows. A brief description of the the methodology to study linguistic leadership by modeling peer influence over a network is in § 1.1, followed by an overview of methods to infer the leaders of semantic change from diachronic text in § 1.2. Each contain the more specific research questions that fall under the broad agenda described so far. This chapter is concluded by stating a concise summary of the thesis and its contributions in § 1.3.

1.1 Sociolinguistic influence on a network

Peer influence in a network is the process in which nodes adopt a behavior knowing that their peers have adopted it. Such imitation of behavior is one reason for node activities to be socially correlated, leading to the production of network cascades. The mechanics of this purely social process are akin to how diseases spread from person to person, and hence peer influence is aptly termed as *social contagion*. Not surprisingly, a variety of social phenomena — ranging from adoption of fashion, communication technology, and memes — have been shown to spread in the network as a result of peer influence (Crane 1999; Yang et al. 2010; Karsai et al. 2014).

Social influence in networks also shapes language change: for example, through the embeddedness of the ties (Labov 1973), the strength or multiplexity of ties (Milroy 1987; Milroy et al. 1992), or the position in the network (Dodsworth 2019). Sociolinguists, having identified the transmission of language change as a fundamental problem to study (Weinreich et al. 1968), have gathered supporting evidence about this fact from relatively small, dense, and self-reported networks. However, quantitative evidence from large behavioral networks, which can additionally show the pathways of influence, has remained elusive both because of a lack of data and tools to validate this claim. Moreover, finer questions about the nature of the ties (e.g., the strength of ties, reciprocity of ties, locality of ties *etc*) and their relation to language change have not been fully answered.

Data-driven models, such as the Hawkes process (HP; Hawkes 1971) models, have been proposed to model influence from different ties in network cascades (e.g., Zhou et al. 2013; Zhao et al. 2015), but they suffer from three issues: first, their parameter space is quadratic in the number of nodes in the network, making them difficult to scale to networks of millions of users; second, the learning procedure typically requires many cascades to reliably estimate the influence parameters; and third, even with scalable modifications (e.g., Li et al. 2014b), these models have not been purposed to test hypotheses about influence, such as if specific ties are more or less influential than others. These issues have made the direct application of these models to test claims about language change difficult.

Aside from sociolinguistics, there has also been a massive historical and contemporary interest in devising methods to detect influence given a network cascade of events in the form of node and timestamp pairs. But many issues make this a challenging task. First, the presence of confounds such as homophily (the tendency of like-minded individuals to get connected) and external stimulus (when some exogenous factor leads to a coordinated behavior) makes this a hard causal inference problem (Shalizi et al. 2011; Shalizi et al. 2016). Second, in many scenarios, network cascades can only be partially observed leading to the possibility of incorrect inferences about influence. Third, the generative process of the cascade, such as the typical temporal spacing of events in a cascade, is not always known in advance, so any method to detect influence should be agnostic to the data generative process. Finally, even if influence is detected, irrespective of the method used, it is also important to characterize the precise role of the underlying social relationships that are more propitious for influence. As relationships between individuals can be articulated differently and networks can be constructed on that basis, it is vital to recognize the relationships that act as conduits for influence from the ones that act as barriers. Experimental methods, despite their success at this task, are not always plausible — for reasons of cost, scale, ethics, or accessibility — so it still remains a goal to detect and characterize influence in purely observational settings.

Despite advances in computational modeling of peer influence, existing methods still remain inadequate to address the aforementioned issues. This part of the thesis uses computational approaches to answer the following research questions about peer influence in network cascades with the aim to close the gap in existing research.

RQ1. Does language change spread on a behavioral network of Twitter mentions and are strong and local ties more likely to lead to adoption of language change? Can HP models be scaled to millions of nodes to answer these questions?

RQ2. Under what conditions can social influence be detected from observational data about event cascades, given the presence of homophily, missing events, and misspecification?

This thesis proposes two modeling ideas that can answer the above questions. The first is to model the temporal dynamics of a cascade as a function of past history, shared network features, and their parameters. This parameterization allows testing the effect of individual features through feature ablation against a goodness-of-fit measure of model performance on the observed cascade. This idea is the basis to answer RQ1. By selecting specific features to proxy for strong and local ties, exploiting the sparsity in the network and the temporal distribution of events in a cascade, the proposed model can be scaled to millions of users on Twitter's social network. The model finds that in adoptions of several non-standard words, peer influence is more effective through a network of densely embedded ties, in comparison to a network of geographically proximate ties.

The second idea is to modify the modeling objective from generating the entire cascade to ranking nodes in the network by their prospect of being the next participant in the cascade, using historical node and dyadic features. This subtle tweak, however, encourages embracing a fully discriminative, Granger-causal approach to detect social influence. Feature ablation can again be utilized to answer RQ2: for example, to test whether peer influence is likely, two competing rankers are designed – one accounts only for homophily and another accounts for both homophily and peer exposure – and their predictions are compared. If prediction improves with the later ranker then peer influence is said to Granger-cause the generation of the cascade; otherwise there is no evidence of peer influence. On synthetic data, in which the conditions of homophily, social influence, missing data, and misspecification is simulated, this discriminative approach is shown to be robust in comparison to existing methods. The effectiveness of this style of testing influence is explored in political finance networks of legislators as well as the propagation of rumors on social media.

1.2 Linguistic leadership for semantic change

Lexical innovations can happen when new wordforms (signs) are introduced or when existing wordforms are repurposed for newer meanings (signifieds) (De Saussure 1916). Words whose meaning change, or semantic innovations, can indicate major linguistic and socio-cultural shifts. As an example, the meaning of the word *gay* changed from being happy to being homosexual, reflecting the cultural shift over the years (Wijaya et al. 2011; Hamilton et al. 2016b). Researchers have contributed many methods to identify semantic innovations from timestamped corpora (see surveys in Kutuzov et al. 2018; Tahmasebi et al. 2018). Others have shown that tracking the semantics of words over time can help in making cultural inferences (Garg et al. 2018; Kozlowski et al. 2019).

Despite the success of these methods, they can only determine what the semantic innovations are from text. By themselves they are inadequate to answer *who* was leading or lagging a particular semantic change. For example, even if a word is identified as a semantic innovation, methods cannot distinguish individual usages of the innovation in a document as leading or lagging *i.e.* if they are used with the contemporary sense of the word or otherwise. Even producers of documents or sources (e.g. authors, newspapers, subreddits, cities, *etc*) can use a semantic innovation with a new meaning before others. Existing methods fail to identify these lead-lag relationships between pairs of sources.

Identifying the leader and follower of an individual semantic change gives a finegrained but isolated view of semantic leadership. Leadership on one change, however, does not necessarily imply leadership on a different change (Raumolin-Brunberg 2006). Without aggregating the changes, it is impossible to reveal whether a source is an overall leader or follower. Aggregation can also reveal the macro-structure of leaders and followers by inducing a leadership network on the sources whose analysis can highlight their roles. Once again, no method has the ability to navigate from reasoning about leadership of individual changes to aggregate leadership over multiple changes, a gap that this thesis aims to fill.

Finally, a relatively recent breakthrough in language research is the use of contextual representations for words in various natural language processing (NLP) tasks (e.g., Peters et al. 2018; Devlin et al. 2019; Raffel et al. 2020). While there have been some attempts to identify semantic changes using contextual word representations (e.g., Giulianelli et al. 2020; Kutuzov et al. 2020), similar to methods that use non-contextual word representations, they are not equipped to identifying leaders of semantic change. One application where this can prove useful is in identifying semantic leaders among publication venues in NLP research such as conferences, journals, workshops that are part of the Association of Computational Linguistics (ACL) anthology, and exploring the connection of linguitic leadership to other explicit notions of influence such as the number of citations.

This part of the thesis asks questions about *who* is leading or lagging semantic changes. The specific research questions are as follows:

RQ3. Given a timestamped collection of documents, can computational methods identify which documents and producers of documents lead with respect to a semantic innovation?

RQ4. Can lead-lag relations for individual semantic innovations be aggregated to induce a semantic leadership network on the producers of documents? Can overall leaders be identified from such a network?

RQ5. Can contextual word representations be used to investigate the connection between semantic leadership and other explicit markers of influence such as the citations?

To answer RQ3, statistical measures based on diachronic word embeddings — dense vector representation of words over time — are used. Semantic innovations are first discovered by comparing the diachronic word embeddings. Then the lead of a document for a semantic innovation is calculated by scoring and aggregating every instance of the innovation in the document for how well the surrounding context can be predicted using the different meanings of the innovation; a usage in the document is said to be "progressive" if the future meaning of the innovation predicts the context better than its past meaning. Qualitative and quantitative evaluation show that this measure of progressiveness of a document indexes important documents in two genres: legal opinions and scientific abstracts. Furthermore, progressiveness of a document is also highly correlated with the documents citation impact in both the domains.

To measure the lead between document sources, a method is proposed to augment temporal embeddings to be conditioned on the source. In effect, the method learns diachronic word embeddings that can further account for the variation in meaning as a result of the source. These specialized embeddings are then used to calculate a statistic by comparing the auto- and cross-correlation between the embeddings for a pair of sources.

To answer RQ4, the method of measuring the lead between document sources is applied on a collection of articles produced by different abolitionist newspapers during the antebellum period and the civil war years. Lead-lag relations between the newspapers for individual semantic changes are filtered by statistically comparisons of their value against randomized data. The retained lead-lag pairs are aggregated to form a network between the newspapers. Centrality analysis on this network using algorithms such as Pagerank (Page et al. 1999) and HITS (Kleinberg 1999) reveals the role by black and women editors in advancing the discourse around the abolition of slavery.

Finally, for RQ5, this thesis proposes a methodological pipeline that combines contextual word embeddings (e.g., BERT; Devlin et al. 2019) with a Hawkes process (HP) model and applies it to the full text of papers from the ACL anthology. Semantic changes and their transition times are first identified; then a classifier is trained to distinguish usages from their “old” sense to “new” sense; usages with the more contemporary sense of any innovation are combined to form event cascades; and finally a HP model is trained to learn the latent influence network between the publication venues of the papers. This setup has multiple advantages: first, it enables the use of the more powerful contextual word embeddings, which, unlike the static embeddings (e.g., Mikolov et al. 2013a; Pennington et al. 2014), are capable of representing multiple co-existing senses to detect semantic changes; second, it uses the probabilistic machinery of HP to induce a network without the need to calculate any lead-lag statistic. This setup shows early promise in investigating the link between the influence of these venues in terms of the citations they get to their linguistic influence.

1.3 Thesis statement and contributions

This thesis can be concisely summarized as follows:

The computational modeling of timestamped text can reveal social dynamics of leadership in language change.

To expand upon the above statement, this thesis makes the following contributions that span multiple research areas through the findings from the different studies.

1. **Sociolinguistic insights about language change on social media**

   This thesis finds that lexical change on social media spreads through the social network in the form of a contagion, highlighting the key role that social influence has in the propagation of language change online. Words follow different contagion characteristics: innovations that are new to the written medium exhibit a complex contagion (Centola et al. 2007), characteristically similar to the propagation of other online innovations (Romero et al. 2011; Karsai et al. 2014; Mønsted et al. 2017), whereas innovations whose origins are from the spoken medium show simple contagion characteristics. Furthermore, linguistic influence in the spread of lexical change is more strongly exerted through densely embedded ties; in contrast, geographically local ties have limited linguistic influence. This contribution complements earlier insights about the spread of language on social media (e.g. Eisenstein et al. 2014; Grieve 2016), and has since spurred interest in deeper computational investigations into other facets of social network structure that affect the spread of lexical change (e.g. Zhu et al. 2021).

2. **Repurposing and scaling event cascade models to detect social influence**

   This thesis proposes methodological extensions to HP models making them suitable to test hypotheses about network structure and language change, at the same time scaling these models to networks of millions of individuals. Both these objectives are achieved by parameter sharing, optimization tricks, and approximations. The diachronic modeling of lexical innovations as event cascades on an underlying social network using these modified HP models reveal insights into the local network mechanism behind diffusion of lexical change on social media; for example, it enables a comparison between different types of ties in the spread of language change. Similarly, a novel discriminative event cascade model is proposed to detect social influence in the presence of homophily and other confounds.

3. **Relation of linguistic influence to other forms of influence**

   This thesis finds evidence that linguistic innovations transmit in the form of a social contagion. This link between social influence on a network driving language change is validated through large-scale network modeling. This thesis also finds many links between linguistic influence or leadership to other forms of influence. In particular, this thesis proposes methods to create diachronic models of semantic change that associate leadership scores to individual documents or between a pair of sources that produce documents. These measures of leadership strongly correlate with influence measured in terms of citations: first, in science and legal domain, documents that are scored highly for their leadership consistently get more citations; and second, venues in computational linguistics that lead other venues on semantic changes are also the also the ones that have more impact. These connections to citation influence are particularly important for the study of science itself (Merton 1973; Fortunato et al. 2018a; Hofstra et al. 2020).

4. **Insights into cultural analytics**

   Language and culture are inseparable (Nida 1998; Jiang 2000). Digital text records — both in historical text archives and from emergent technologies such as social media — encode our cultural history. This thesis proposes computational methods and techniques that can ingest data at scale to summarize what is changing and who is leading changes, which makes it possible to gain insights in the analysis of culture. In the analysis of historical text collections from the legal domain, this thesis can index landmark court opinions based on whether they are at the forefront of semantic change: for example, court opinions that are taken as precedent on issues of abortion are found by methods described in this thesis to be semantically innovative with respect to terms such as *fertilization*. Similarly, in the analysis of abolitionist newspapers from the nineteenth century, the computational methods proposed in this thesis highlight the role of editors from minority groups in the abolition of slavery.

CHAPTER 2
BACKGROUND

This chapter reviews the relevant prior work for this thesis. An overview of past work on language change is presented first in § 2.1. This section presents a brief historical account of research on language change but emphasizes the more recent and computational research work on language change that is relevant to situate the linguistic contributions of this thesis. Next, in § 2.2 a brief summary of related methods is presented. This section covers key background about computational methods to model the diffusion dynamics on a network (§ 2.2.1), and language processing methods to detect lexical semantic change from text (§ 2.2.2). This section will help in situating the methodological contributions of this thesis. Finally, this thesis makes contributions to to other research disciplines, which are drawn out with the background work in those areas in § 2.3.

2.1 Language change

The study of language change has a long history in sociolinguistics. Sociolinguists seek to understand both the fine structure of language but also the social dimension by studying variation and change. For example, it is not merely of interest to know that the pronoun *thou* was abandoned in favor of *you* but the motivation behind the change uncover the regional and social differences that led to the change (Trudgill 2009). To this end, sociolinguists have established strong connection of language variation and change to a variety of social factors, such as the construction and evaluation of social identities (e.g., Bucholtz 1999a; Bucholtz et al. 2005); individual attributes such as race (e.g., Bucholtz 1999b), gender (e.g., Eckert et al. 2013); constructs such as social class (e.g., Labov 1990); all the way to differences in the organizational structure based on geography (e.g., Trudgill 1974) and social networks (e.g., Milroy et al. 1985; Kerswill et al. 2000).

Much of the traditional sociolinguistic work on language change, however, differs with the work in this thesis along three salient aspects. First, the traditional studies operate on limited data either elicited from interviews (e.g., D'Arcy et al. 2015) or from observations within a speech community (e.g., Eckert 1989). Second, they rely heavily on the method of *apparent time* (Bailey 2002), in which changes in progress are identified by comparing the speech of younger and older individuals. Older individuals are considered to lag changes compared to younger individuals who are considered the drivers of language change. The key assumption is that each person's use of language is stable throughout their adult life, a hypothesis that has been broadly supported by prior work on changes in phonology and grammar (Sankoff et al. 2007). Third, the studies are almost exclusively focused on sound (e.g., Labov 1963) and morphosyntactic change (e.g., Eckert et al. 2013; Tagliamonte et al. 2009).

In contrast, this thesis develops methods and bases its findings from written timestamped text. This difference, in conjunction with using computational methods, helps scale to large amount of language data and potentially study more and varied changes in language. Moreover, instead of comparing the rate of variation between speakers of different age groups, changes are identified by comparing usage across time, or as is technically called *real time* analysis. This difference is key as it enables identifying and studying accelerated language changes that may affect the language in a single generation (Dannenberg 2000). Finally, this thesis studies changes in language at the level of words — both in the introduction and transmission of non-standard terms in a network, and the shift in the meaning of words and their propagation on a latent network. While the goal of studying transmission of language in the network is not new, it has not been studied in the context of semantic changes. Sociolinguists have long been cognizant of the opportunity to study semantic changes but lacked the methodological tools for detecting and analyzing their spread in the network (see discussion in Labov et al. 1973), a gap that this thesis fills.

The core question of this thesis concerns with finding linguistic leaders of change from diachronic text data. This is considered by studying the dynamics of transmission of change, a problem of fundamental importance to sociolinguistics (Weinreich et al. 1968), and in identifying the leaders of lexical semantic change. The relevant prior work for both is discussed next.

2.1.1 Language change and social networks

Social networks have been considered important for language change (e.g., Labov 2001; Milroy et al. 2013; Bot et al. 2002). Prior work in sociolinguistics has linked the network characteristics of individuals with their positioning with respect to language change: for example, Labov (2001) argues that individuals with many weak ties tend to be leaders, Milroy (1987) finds that dense sub-networks of strong ties are less welcoming of change, analysis of real as well as synthetic networks also points to the importance of network position to language change (e.g., Fagyal et al. 2010; Dodsworth 2019).

Traditional sociolinguistic approach to connect social networks and language change is restricted by the small scale of these networks typically because they are obtained by snowball sampling. Another important limitation is the overreliance on ego-networks or local sociograms, in part due to difficulty in collecting and analyzing social network data, which restricts inferences to individual or properties of individuals in a network (Sharma et al. 2020). This thesis, on the other hand, uses large networks, going beyond just the ego-alter links, that makes it possible to connect individual, dyadic, and network structural characteristics to language change.

More recently, the computational research in sociolinguistics has investigated the network aspect of language change on social media. On Twitter, Eisenstein et al. (2014) construct a network of sociolinguistic influence between cities in the United States. Similarly, Grieve et al. (2018) identified the hubs and pathways of diffusion across cities for the lexical innovations that emerged on Twitter, though without explicitly constructing a network. This thesis shares similarities with these works but deviates in two ways: first, for lexical innovations the thesis aims to model the influence of structural characteristics of the network of Twitter users; second, the thesis extends the notion of inducing a network to lexical semantic changes.

Outside of sociolinguistics, some attempts to quantify network influence in text corpora have been made. Almost exclusively, these approaches operate at the level of change in individual word frequencies or change in topics: for example, Guo et al. (2015) measure influence within small groups by the use of individual words, on the assumption that Alice is likely to have influenced Bob if words used by Alice tend to be subsequently used by Bob. Similarly, text reuse has been used as a signal to induce an influence network between newspapers (Smith et al. 2013). Topic models (Blei 2012) have also been used to measure influence in networks (Tang et al. 2009; Liu et al. 2010). In contrast to these works, a large part of this thesis focuses on quantifying network influence for lexical semantic changes, a topic that has not been explored sufficiently. While lexical frequencies can be informative — particularly in cases that fit theoretical models like communicative accomodation theory (Giles et al. 1991) — they are often caused by outside events (e.g., new people or keywords) rather than conceptual changes in the discourse which are captured by semantic changes.

2.1.2 Lexical semantic change and leadership

The dynamics of change in a language are extended to the entire lexicon, which links words and their semantics (Pierrehumbert 2012). One of the two types of language change considered in this thesis is lexical semantic change — when meaning of words shift over time (Traugott et al. 2001). These changes are ubiquitous: they are observed in different languages (Hamilton et al. 2016b); at different timescales (e.g., *gay* changed over decades in the twentieth century (Kulkarni et al. 2015) whereas *lit* has gained an additional slang sense in a matter of years (Shoemark et al. 2019)); and in different styles (e.g., the skull emoji in online writing; Robertson et al. 2021).

This linguistic phenomena has been studied qualitatively and through philological methods (e.g., Zalizniak et al. 2012; Geeraerts 1990). Early quantitative methods to detect semantic changes focused on operationalizing the change through frequency statistics of words (e.g., Bybee 2007). The more recent advance in semantic change detection is due to models that represent meaning of words as vectors in space (Mikolov et al. 2013a; Pennington et al. 2014; Baroni et al. 2014). This idea has further been extended to diachronic text where many methods learn to represent word meaning as a sequence of vectors over time (e.g., Hamilton et al. 2016b), and semantic changes are detected by comparing the differences in these vectors (e.g., Hamilton et al. 2016a). A more thorough review of semantic change detection methods is in § 2.2.2.

The utility of temporal word vectors has been shown both to gain linguistic and social insights. Hamilton et al. (2016) treat temporal word vectors as evidence for structural constraints on semantic change, while Garg et al. (2018) trace shifts in public attitudes towards race and ethnicity by tracking timestamped word vectors in large-scale corpora of books. In the digital humanities, temporal word vectors have been used to trace the history of concepts: for example, Shechtman (2020) explores the "technical, ideological, and environmental valences" of the idea of media by comparing near neighbors of *media* and related terms across a corpus of 20th century magazines.

Much of the focus of the aforementioned research, is in finding semantic changes when when these changes are used as artefacts that represent broader changes in society or culture. Despite an array of methods to detect semantic changes, very few go the extra yard to find the leaders of semantic changes. Methodologically, this thesis diverges from this prior work by focusing on identifying leaders of individual semantic changes, and then computing and analyzing aggregate leadership networks over hundreds of changes. This thesis also proposes the combination of network and contemporary computational linguistic methods, such as contextual word embeddings (e.g., BERT; Devlin et al. 2019), to learn the semantic leadership network from timestamped text. Finally, this thesis demonstrates the links between lexical semantic leadership to broader notions of influence.

2.2 Methodological foundations

Throughout the thesis, a network perspective to model the propagation of linguistic innovations is presented with the objective of identifying the leaders and followers. These models are applied to two types of changes: lexical and lexical semantic change. Identifying the latter type of change requires the application of computational methods to large volumes of language data. Here, a brief background of network propagation models and semantic change detection models is provided in relation to the methodological contributions of this thesis.

2.2.1 Network cascade models

Research in the diffusion of innovations has an extensive history (Rogers 2010), including more specifically on diffusion in a network (Valente 1996). Diffusion models have been used to understand the propagation processes in a variety of social and information networks: for example, in finding influential nodes in the network that maximize diffusion (e.g., Kempe et al. 2003; Aral et al. 2012); in the propagation of information (e.g., Adar et al. 2005; Romero et al. 2011; Leskovec et al. 2007b; Gomez-Rodriguez et al. 2014); in the adoption of new products (e.g., Watts et al. 2007; Aral et al. 2011a); in viral marketing (e.g., Leskovec et al. 2007a; Lappas et al. 2010), *etc*.

Recently, cascades of events (e.g., adoption of a new product or a word) on a network have been modeled as point processes (e.g., Du et al. 2016; Upadhyay et al. 2018; Mei et al. 2017; Zuo et al. 2020);[1] particularly popular are a family of parametric inhomogeneous Poisson processes called as the Hawkes process (HP; Hawkes 1971). The key idea in HP is that the rate and timing of events in the future depend on the entire history of past events; in multivariate HP, the timing and the source of the event depends on the entire history across

[1] A more comprehensive and theoretical review of point processes is in Daley et al. (2003)

all sources. HP models have attractive features: first, the influence from every individual event in the past to the future is considered in modeling; second, the influence — even from different sources — can be aggregated and adjusted in a probabilistic learning framework with the choice of temporal kernels and model parameters.[2] HP models and their extensions have been used successfully to forecast the size of network cascades (e.g., Zhao et al. 2015) and to recover the latent network structure from event cascades (e.g., Zhou et al. 2013).

However, estimating the number of parameters in multivariate HP can become quadratic in the number of nodes in the network making inference computationally intractable; the learning also comes at the cost of computing complicated integrals that limit scalability to large-scale networks. The basic formulation of HP is not suitable to test hypotheses about the structural properties of the network in the spreading processes. This thesis uses HP models and proposes solutions to overcome these disadvantages (see chapter 3 and chapter 4).

2.2.2 Semantic change detection models

Aside from network propagation models, computational models to detect semantic change from text are also a focus of deeper investigation throughout this thesis. Though early computational approaches to detect semantic changes were either based on frequency (e.g., Clarke et al. 1991; Bybee 2007) or topic models (e.g., Sagi et al. 2011), the more contemporary methods are based on word embeddings — vector representations of words learned from text. Representing word meaning as vectors in space is in keeping with the distributional hypothesis in language (Harris 1954; Firth 1957), which asserts that the meaning of a linguistic element can be ascertained by the contexts in which it appears; the vector space models learn to map words to vectors by accounting the contexts in which words appear, such that words with similar meanings are closer in vector space.

This idea has further been extended to diachronic text where many methods learn to

[2] The subsequent chapters review HP models more thoroughly but a historical and mathematical review of HP models can be found in Rasmussen (2011)

represent word meaning as a sequence of vectors — called diachronic word embeddings — evolving over time (Cook et al. 2010; Jatowt et al. 2014; Gulordava et al. 2011; Kim et al. 2014; Kulkarni et al. 2015; Hamilton et al. 2016b; Frermann et al. 2016; Rudolph et al. 2018). A typical approach is to estimate embeddings on multiple corpora from different time periods and then align the embedding vectors to make them comparable; however, there are many alternatives (for an overview, see Kutuzov et al. 2018; Tahmasebi et al. 2018; Tang 2018). Semantic changes are then detected using diachronic embeddings by measuring the differences in the position of the vectors or through auxiliary measures that consider the neighboring vectors (Hamilton et al. 2016a; Schlechtweg et al. 2019; Gonen et al. 2020).

While most semantic change detection methods rely on static embeddings — that is, a single vector per word — to represent meaning, the recent introduction of contextual language models (e.g., Peters et al. 2018; Devlin et al. 2019), which assign a vector to every single usage of a word, has also been explored by some methods to detect semantic changes (e.g., Giulianelli et al. 2020). Though the effectiveness of using contextual word embeddings to detect semantic changes is still an open question (Kaiser et al. 2020; Laicher et al. 2020), an advantage they offer over static embeddings for change detection is the potential to tease apart co-evolving senses of the same word over time. In chapter 7 of this thesis, contextual word embeddings are used to detect semantic changes in papers that are published in the ACL anthology.

A longstanding impediment in semantic change detection is the lack of groundtruth for systematically evaluating the semantic change models. Researchers have resorted to creative evaluations that include, checking if the models pick out words that are known to have changed (e.g., Kulkarni et al. 2015); comparing the changes from a model to human judgments (Del Tredici et al. 2019); predicting document timestamps using the changes as features (Mihalcea et al. 2012); synthetically injecting changes in a corpus that the model picks out (Shoemark et al. 2019); and through hypothesis testing (Hamilton et al. 2016b).

More recently, researchers have started annotating semantic changes in well-known corpora which has the advantage of a precision-recall style evaluations (Schlechtweg et al. 2018; Schlechtweg et al. 2020); however, a limitation is that the changes are specific to the corpora that was used for annotation. This thesis takes the approach of evaluating the utility of semantic changes by using them to test hypotheses about leadership.

2.3 Intersections with other research areas

2.3.1 Computational social science

One of the most recognized interdisciplinary research areas where computational methods have made a big impact is computational social science (Lazer et al. 2009). Computational social science as a disciplinary area involves a rich variety of data, tools, and methods to study social phenomena. The past decade has seen rapid advancement in computational linguistic methods which has also coincided with a huge surge of text data produced by users who use the digital medium, but also in the mass digitization of historical archives. *Text as data* is an emerging paradigm in which computational methods are applied on digital texts to make sociocultural inferences (Gentzkow et al. 2019). The increase in the availability of digitized text coupled with advances in computational methods has ushered in a new era in which this paradigm has flourished (Salganik 2019). Recent works have demonstrated the promise of this paradigm across a range of phenomena - for example, in inferring power relationships between interlocutors (Danescu-Niculescu-Mizil et al. 2012; Prabhakaran et al. 2012), estimating political slant from text (Gentzkow et al. 2019; Sim et al. 2013), pointing racial disparities in conversations between law enforcement officers and community members (Voigt et al. 2017), and many others.

There is also a long line of work in applying the text-as-data paradigm to timestamped text. This comprises of works in computational sociolinguistics towards studying the underlying factors of a change (Danescu-Niculescu-Mizil et al. 2013; Eisenstein et al. 2014); developing methods to model the changes in text (Blei et al. 2006; Wang et al. 2006;

Kulkarni et al. 2015; Hamilton et al. 2016b); and using text to make sociocultural inferences (Golder et al. 2011; Underwood et al. 2018; Garg et al. 2018). Two particularly relevant lines of work in computational social science that are relevant for this thesis are modeling the diffusion in social networks (e.g. Bakshy et al. 2012; Cheng et al. 2016) and the study of science (e.g. Fortunato et al. 2018a)

This thesis makes contributions to the field of computational social science. chapter 3 describes scalable computational models for diffusion of language change on social networks that are fully equipped to test network hypothesis about change. Though applied in the setting of language change, similar ideas from modeling can be used for general problems such as detection of social influence in event cascades that are relevant in other applications(see chapter 4). On the other hand, chapter 5 and chapter 7 describe methods that establish previously unexplored links between semantic innovativeness to other proxy measures of influence in science, such as the number of citations.

2.3.2 Cultural analytics

Over the years, cultural analytics has shaped up as a new highly encompassing research area. Cultural analytics encourages the use of computational methods to operate on cultural data at scale (Manovich 2020). A key objective of the field is to take a “big data” approach to study cultural forms with the view that this can bring out the variation and diversity in both the cultural artefacts and the processes that produce them (Manovich 2016). Based on its stated objectives and the emphasis on using data-driven methods, cultural analytics has unified seemingly disparate areas such as social computing and digital humanities.

Text is one source of social and cultural data (Nguyen 2017; Nguyen et al. 2020). As digitization has increased, research showcasing longitudinal analyses of text to understand the evolution of culture have increased hand in hand (e.g. Michel et al. 2011; Leetaru 2011, i.a). This thesis makes contributions to the field of cultural analytics. Methods that cultural analysts can find useful irrespective of the type of text they want to understand are proposed throughout the thesis. In chapter 5, a scheme is proposed to score a unit of text (e.g. sentence, paragraph, document, chapters, books, *etc*) for its semantic progressiveness — that is, the extent to which one or more words are used with a meaning before that meaning became dominant. Such a method can potentially offer a text analytic tool for a humanist to navigate the spectrum of distant to close reading (Jänicke et al. 2015). In chapter 6, a method that goes from a collection of timestamped documents produced by different sources to a network of leadership on semantic innovations is shown. On a set of abolitionist newspapers, which are an important historical and cultural artefact, this method brings to attention the importance of editors from minority groups that have not been given enough credit for their role in the abolition of slavery.

CHAPTER 3
NETWORK DYNAMICS OF LEXICAL CHANGE

3.1 Motivation

Language change is fundamentally a social phenomenon (Labov 2001). For newer linguistic forms to become widespread, two conditions need to be met: first, speakers (and writers) should come into contact with the new form; second, they must decide to use it. The first condition implies that language change is related to social network structure (Milroy et al. 1992), since the exposure of linguistic variation often comes from network ties. But while the first condition is necessary, it is by no means sufficient. The second condition implies that even after exposure individuals make strategic choices, often to construct and evaluate their own social identity, when deciding whether to adopt new linguistic forms (Bucholtz et al. 2005; Johnstone et al. 2002; Labov 1963). The analysis of language change patterns can thus uncover the latent structure of social organization: to whom people talk, from whom they are influenced, and how they see themselves.

Though the basic outline of the interaction between language change and social structure is known, the fine details are still missing: What local mechanism explains the overall spread of language change in a large social network? What types of social network connections are most important for language change? To what extent do considerations of identity affect linguistic differences, particularly in an online context? Traditional sociolinguistic approaches lack both the data and the methods for asking such detailed questions about language variation and change.

This work shows the use of large-scale social media data to answer the above questions. A large, temporally distributed data set of Twitter users that contains all public messages for several million accounts, and augmented with social network and geolocation metadata, is collected with the aim to tracking, and potentially explaining, every usage of a linguistic variable[1] as it spreads through social media.

Citation Rahul Goel, Sandeep Soni, Naman Goyal, John Paparrizos, Hanna Wallach, Fernando Diaz, and Jacob Eisenstein. "The social dynamics of language change in online networks". In: *International Conference on Social Informatics*. Springer. 2016, pp. 41–57

3.2 Hypotheses

This work tests three main hypotheses related to the transmission of linguistic variables. Our variables are non-standard words with strong and well-known geographic associations. The details on how these variables were selected are in § 3.3.

H1 Non-standard words propagate between individuals who are connected in the Twitter social network in the form of a contagion.

H1 implies that language change can be viewed as social influence or a contagion, one in which the adoption likelihood of linguistic markers depends on past exposures from network ties. Qualitative evidence in support of H1 is limited to networks of few individuals (Milroy et al. 1992). This work tests H1, however, at a scale of millions of users in examining the link between peer influence and language change in online networks.

The next two hypotheses test whether specific types of social network connections are more influential than others. The sociolinguistic theory of local *covert prestige*—i.e., the positive social evaluation of non-standard dialects affects the adoption of new linguistic forms—suggests that geographically local ties are more influential than non-local ties (Trudgill 1972). But quantitatively validating the theory has been difficult, in part because many explanations fit the observation that people often use similar language to that of their geographical neighbors. One is exposure: due to geographical assortativity in online social ties (Al Zamal et al. 2012), people are likely to be exposed to local linguistic forms

[1] Sociolinguistic and dialectogical literature refer to a basic unit of linguistic differentiation as a "variable" (Wolfram 1991)

through local ties. Alternatively, language similarity can be the reason for individuals to prefer social ties with other local individuals.[2] Here, the aim is to test the following hypothesis about the influence of local ties in language change using a model that makes it possible to tease apart the roles of geographic assortativity and local influence.

H2 Local ties are better conduits for transmission of non-standard words.

Similarly network connections can be differentiated in terms of strong and weak ties (Granovetter 1973; Milroy et al. 1992). Strong ties signify more important social relationships, and are often densely embedded, meaning that the nodes in question share many mutual friends; in contrast, weak ties act as bridges between disconnected communities. Past research investigating information diffusion, through resharing of URLs on Facebook, found weak ties to play an important role: URLs shared across strong ties were likely to be reshared, but the abundance of weak ties and the higher chances of them being a source of novel information made propagation through weak ties more likely (Bakshy et al. 2012). In some respects, language change is similar to traditional information diffusion scenarios, such as resharing of URLs. But, in contrast, language connects with personal identity on a much deeper level than a typical URL. As a result, strong, deeply embedded ties may play a more prominent role in the adoption of language change markers. This motivates to test the following hypothesis.

H3 Strong or densely embedded ties are better conduits for transmission of non-standard words.

3.3 Data

The dataset for this study is acquired via an agreement between Microsoft and Twitter. It contains all public messages between June 2013 and June 2014 by several million users.

[2] This is an instantiation of a general problem about detecting social influence in cascading behavior that it is often confounded with homophily (Shalizi et al. 2011; Shalizi et al. 2016); a modeling solution to this more general problem is the focus of chapter 4.

Additional metadata for these users — the social network and geolocation — is also included. Any retweets, non-English tweets and tweets from outside the United States are excluded during preprocessing. The acquisition of this dataset is aimed to overcome the main limitation of selection biases that persist in small-scale network studies of language change using traditional methods.

3.3.1 Linguistic variables

There is overwhelming evidence that written language on social media continues to change at all levels, from spelling all the way up to syntax structure and semantic meaning across the lexicon (Eisenstein 2015a; Tagliamonte et al. 2008; Del Tredici et al. 2017). But the focus for this study is on words, which are strategically chosen as the linguistic variables because they are most easily observable and measurable using computational methods. A set of non-standard words with a strong geographical association to eight cities in the United States—representing a diverse mix of geographic regions, population densities and demographics—are taken as the linguistic variables. These variables are identified using SAGE (Eisenstein et al. 2011), following the approach described in (Eisenstein 2015b).[3] The words and their associated cities are shown in Table 3.1.

The words are also divided into three main classes: (a) **Lexical words** (e.g. *cookout*, *graffiti*, *jawn*, *hella*[4], *phony*, and *stamp*), whose origin can almost certainly be traced back to spoken language; (b) **Phonetic spellings** (e.g. *ain*, *ard*, *asl*, *dese*, *inna*, and *yeen*), which are non-standard spellings based on phonetic variation by region, demographics, or situation; and (c) **Abbreviations** (*ctfuu*, *lls*, *mfs*, and *tfti*), which are intrinsic to social media writing.

[3] After running SAGE to identify words with coefficients above 2.0, a manual step was taken to remove hashtags, named entities, non-English words, and descriptions of events.

[4] *hella* (Bucholtz et al. 2007), and *jawn* (Alim 2009) are dialect markers that are not widely used outside their regions of origin, even after several decades of use in spoken language.

Table 3.1: **Linguistic variables and their associated cities.** The words are distributed across 8 major metropolitan areas. The meaning of some words that are not widely known are provided in parentheses.

City	Linguistic variables
Atlanta	*ain* (phonetic spelling of *ain't*), *dese*(phonetic spelling of *these*), *yeen* (phonetic spelling of *you ain't*)
Baltimore	*ard* (phonetic spelling of *alright*), *inna* (phonetic spelling of *in a* and *in the*), *lls* (*laughing like shit*), *phony* (fake)
Charlotte	*cookout*
Chicago	*asl* (phonetic spelling of *as hell*, typically used as an intensifier on Twitter), *mfs* (*motherfuckers*)
Los Angeles	*graffiti*, *tfti* (*thanks for the information*)
Philadelphia	*ard* (phonetic spelling of *alright*), *ctfuu* (expressive lengthening of *ctfu*, an abbreviation of *cracking the fuck up*), *jawn* (generic noun)
San Francisco	*hella* (an intensifier)
Washington D.C.	*inna* (phonetic spelling of *in a* and *in the*), *lls* (*laughing like shit*), *stamp* (an exclamation indicating emphasis).

3.3.2 Social Network

The social network between the Twitter users is constructed from their mutual replies. Specifically, every Twitter user is represented as a node in a graph and an undirected edge between a pair of users is placed if each replied to the other. The decision to use the reply network (rather than the follower network) is both a pragmatic and an informed choice: the follower network is not widely available and previous research from Huberman *et al.* argues that Twitter's mention network is more socially meaningful than its follower network (Huberman et al. 2008). The focus on mutual replies alleviates the problem of a large number of unrequited replies directed at celebrities. The network characteristics of the undirected social network constructed from mutual replies is compared to the characteristics of a directed social network of replies in Figure 3.1.

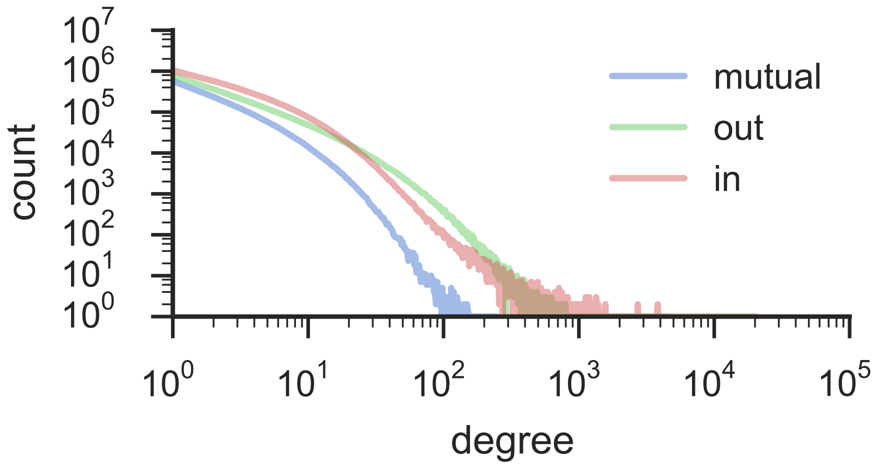


Figure 3.1: **Degree distributions.** The degree distribution for the mutual-reply network and the in and out networks shown on log-log scale. Visually, the network shows a power law distribution for the degrees.

3.3.3 Geography

The geographical information is obtained from the metadata from Twitter's location field, populated via a combination of self reports and GPS tagging. The metadata is aggregated across each user's messages, to geolocate each user to the city from which they most commonly post. The data set contains 4.35 million geolocated users, of which $589,562$ were geolocated to one of the eight cities listed in Table 3.1.

3.4 Methods

A combination of statistical testing and data-driven modeling is used to test all the hypotheses in $\S$ 3.2.

Validating `H1` This involves estimating the fraction of users who used one of the words listed in Table 3.1 only after being exposed to that word by one of their social network connections. Formally, user i EXPOSES user j to word w at time t if and only if the following conditions hold: i used w at time t; j had not used w before time t; the social network connection $i \leftrightarrow j$ was formed before time t. The individual exposures are aggregated across

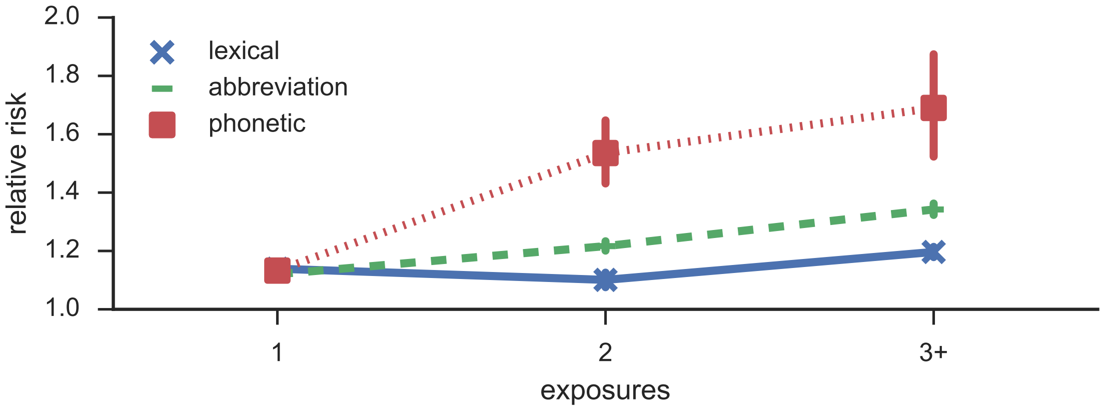


Figure 3.2: **Infection risk as a function of number of exposures.** The relative infection risks for words in each of the three linguistic classes defined in § 3.3. The figure depicts 95% confidence intervals, computed using the shuffle test (Anagnostopoulos et al. 2008).

all users in the network to calculate INFECTION RISK for word w, defined as the number of users who use word w after being exposed divided by the total number of users who were exposed. To consider the possibility that multiple exposures have a greater impact on the infection risk, the infection risk after exposures across one, two, and three or more distinct social network connections is also calculated.

The infection risk of a word cannot be interpreted directly as evidence for social contagion: since infection risk is a measure of social correlation, high infection risk can also be explained by homophily or external confounds. To address this, the SHUFFLE test is used, which compares the observed infection risk for a word to infection risks under the null hypothesis that event timestamps are independent (Anagnostopoulos et al. 2008). The null hypothesis infection risks are computed by randomly permuting the order of word usage events. If the observed infection risks are substantially higher than the infection risks computed using the permuted data, then this is compatible with social influence.

Figure 3.2 depicts the ratios between the words' observed infection risks and the words' infection risks under the null hypothesis, after exposures across one, two, and three or more distinct connections. For all the linguistic classes defined in § 3.3, the risk ratio for even a single exposure is significantly greater than one, suggesting the existence of social influence. The risk ratio for a single exposure is nearly identical across the three classes. For phonetic spellings and abbreviations, the risk ratio grows with the number of exposures, suggesting that words in these classes exhibit COMPLEX CONTAGION—i.e., multiple exposures increase the likelihood of adoption (Centola et al. 2007). In contrast, the risk ratio for lexical words remains the same as the number of exposures increases, suggesting that these words spread by simple contagion.

For linguistic markers that have not undergone widespread adoption, such as phonetic spellings and abbreviations, the adopters risk negative social evaluations of their linguistic competency, as well as their cultural authenticity (Squires 2010). In contrast, lexical words are already well known from spoken language and thus carry low social risk. This difference may explain why lexical words do not exhibit complex contagion.

Validating `H2` and `H3` Testing `H2` and `H3` requires a more sophisticated modeling tool than simply estimating exposure counts. To compare the impact of exposures across different types of social network connections requires asking *who* exposed the user to the linguistic marker. Moreover, exposures need to be credited considering their temporal properties—for example, exposure from a weak tie in the past hour should be treated differently to an exposure from a strong tie in the past day. This motivates a probabilistic modeling approach, in which the Twitter data set is seen as a set of cascades of timestamped events, one per geographically distinctive word described in Table 3.1. Each event in a word's cascade corresponds to a tweet containing that word, and each cascade is modeled as a probabilistic process whose parameters are learned.

The core modeling framework is based on a HAWKES PROCESS (HP; Hawkes 1971)—a specialization of an inhomogeneous Poisson process—which explains a cascade of timestamped events in terms of influence parameters. In a temporal setting, an inhomogeneous Poisson process says that the number of events y_{t_1,t_2} between t_1 and t_2 is drawn from a Poisson distribution, whose parameter is the area under a time-varying INTENSITY FUNCTION over the interval defined by t_1 and t_2:

$$y_{t_1,t_2} \sim \text{Poisson}\left(\Lambda(t_1,t_2))\right), \tag{3.1}$$

where $\Lambda(t_1,t_2) = \int_{t_1}^{t_2} \lambda(t) \, \mathrm{d}t$. Since the parameter of a Poisson distribution must be non-negative, the intensity function must be constrained to be non-negative for all possible values of t.

An HP is a self-exciting inhomogeneous Poisson process, where the intensity function depends on previous events. For a cascade of N events $\{t_n\}_{n=1}^{N}$, where t_n is the timestamp of event n, the intensity function is

$$\lambda(t) = \mu_t + \sum_{t_n \leq t} \alpha \cdot \kappa(t - t_n), \tag{3.2}$$

where μ_t is the base intensity at time t, α is a parameter that captures the influence of previous events, and $\kappa(\cdot)$ is a time-decay kernel to adjust the influence of previous events. Though there can be many choices for the temporal decay kernel, an exponential decay kernel, which ensures that events in the distant past carry much less influence, is the typical choice in many applications.

The framework can be extended to vector observations $\boldsymbol{y}_{t_1,t_2} = (y_{t_1,t_2}^{(1)}, \ldots, y_{t_1,t_2}^{(M)})$ and intensity functions $\boldsymbol{\lambda}(t) = (\lambda^{(1)}(t), \ldots, \lambda^{(M)}(t))$, where, in this setting, M is the total number of users in the data set. For a cascade of N events $\{(t_n, m_n)\}_{n=1}^{N}$, where t_n is the timestamp of event n and $m_n \in \{1, \ldots, M\}$ is the source of event n, the intensity function for user $m' \in \{1, \ldots, M\}$ is

$$\lambda^{(m')}(t) = \mu_t^{(m')} + \sum_{t_n < t} \alpha_{m_n \to m'} \cdot \kappa(t - t_n), \tag{3.3}$$

where $\mu_t^{(m')}$ is the base intensity for user m' at time t, $\alpha_{m_n \to m'}$ is a pairwise influence parameter that captures the influence of user m_n on user m', and $\kappa(\cdot)$ is a time-decay kernel. In the experiments, the exponential decay kernel is $\kappa(\Delta t) = e^{-\gamma \Delta t}$ and γ is set so that $\kappa(1 \text{ hour}) = e^{-1}$.

Researchers usually estimate all M^2 influence parameters of HP (e.g. Li et al. 2014a; Zhao et al. 2015). However, when $M > 10^6$, as is the case for this study, estimating $O(10^{12})$ influence parameters is computationally and statistically intractable, given that our data set includes only $O(10^5)$ events. More critically, directly estimating these parameters does not help in quantitatively testing the social evaluation hypotheses. Instead, Li and Zha's parametric HP model is utilized (Li et al. 2014b). This model defines each pairwise influence parameter as a linear combination of pairwise features:

$$\alpha_{m \to m'} = \boldsymbol{\theta}^\top \boldsymbol{f}(m \to m'), \tag{3.4}$$

where $\boldsymbol{f}(m \to m')$ is a vector of features that describe the relationship between users m and m'. Thus, one only needs to estimate the feature weights $\boldsymbol{\theta}$ and the base intensities. This modification drastically simplifies estimation, enables the model to share statistical strength across dyads, and scale the framework to millions of users. The reduction in parameters can take away some of the predictive power of the model but equips the model to compare explanations, a tradeoff that is favorable to this study. Since the timespan considered is only for an year, as a further simplification, all the base intensities are assumed to be independent of time.

To ensure that the intensity functions $\lambda^{(1)}(t), \ldots, \lambda^{(M)}(t)$ are non-negative, $\boldsymbol{\theta}$ and the base intensities are assumed to be non-negative. The parameters are learned through coordinate descent by maximizing the constrained likelihood of the cascade data. To increase efficiency, several approximations that exploit the sparsity in the temporal distribution of the cascades are made to scale the estimation procedure to millions of users. The details about optimization and various approximations are in appendix A.

3.5 Experimental setup

Four binary features are designed to enable testing the hypotheses about the roles of different types of social network connections:

F1 Self-activation: This feature fires when $m' = m$. It captures the scenario where using a word once makes a user more likely to use it again, perhaps because they are adopting a non-standard style.

F2 Mutual reply: This feature fires if the dyad (m, m') is in the Twitter mutual-reply network described in § 3.3. This feature is also used to define the remaining two features, which ensures that features F2, F3, and F4 were (at least) as sparse as the mutual-reply network.

F3 Tie strength: This feature fires if the dyad (m, m') is in in the Twitter mutual-reply network, and the Adamic-Adar value for this dyad, a measure of tie strength (Adamic et al. 2003), is especially high — *i.e.*, above the 90th percentile among all dyads where at least one user has used the word in question. Thus, this feature picks out the most densely embedded ties.

F4 Local: This feature fires if the dyad (m, m') is in the Twitter mutual-reply network, and the users were geolocated to the same city, and that city is one of the eight cities listed in Table 3.1. For other dyads, this feature returns zero. Thus, this feature picks out a subset of the geographically local ties.

A separate set of parametric HP models for each of the geographically distinctive linguistic markers is used. Specifically, for each word, a set of nested models is constructed by first creating a baseline model using features F1 (self-activation) and F2 (mutual reply) and then adding in each of the experimental features—*i.e.*, F3 (tie strength) and F4 (local). The hypothesis `H3` (strong ties are more influential) is tested by comparing the goodness of fit for feature set F1+F2+F3 to that of feature set F1+F2. Similarly, `H2` (geographically

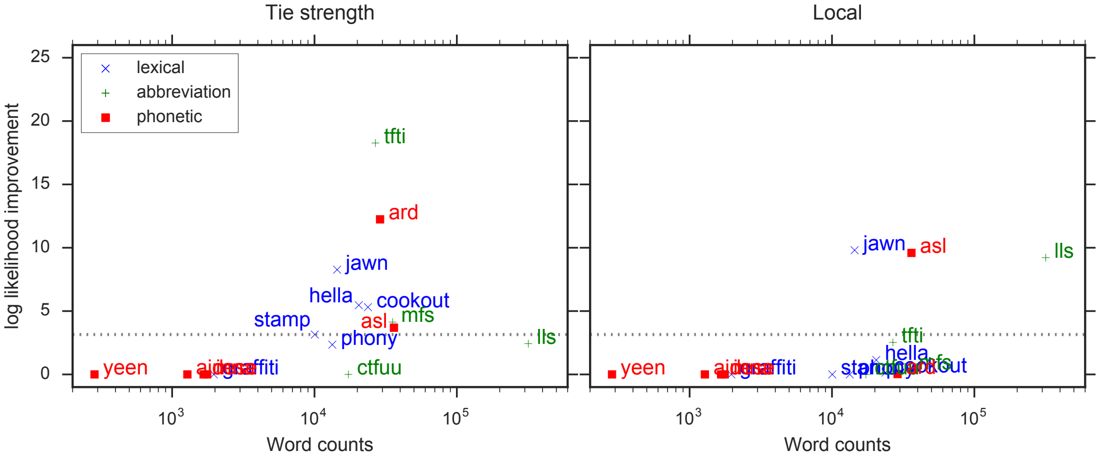


Figure 3.3: **Hypothesis testing results.** Improvement in goodness of fit from adding in features F3 (tie strength) and F4 (local). The dotted line corresponds to the threshold for statistical significance at $p < 0.05$ using a likelihood ratio test with the Benjamini-Hochberg correction (Benjamini et al. 1995).

local ties are more influential) is tested by comparing the goodness of fit for feature set F1+F2+F4 to that of feature set F1+F2.

3.6 Results

Figure 3.3 shows the improvement in goodness of fit from adding in features F3 and F4. Under the null hypothesis, the log of the likelihood ratio follows a χ^2 distribution with one degree of freedom, because the models differ by one parameter. Features F3 and F4 did not improve the goodness of fit for less frequent words, such as *ain*, *graffiti*, and *yeen*, which occur fewer than 10^4 times. Below this count threshold, there is not enough data to statistically distinguish between different types of social network connections. However, above this count threshold, adding in F3 (tie strength) yielded a statistically significant increase in goodness of fit for *ard*, *asl*, *cookout*, *hella*, *jawn*, *mfs*, and *tfti*. This finding provides evidence in H3's favor—that the linguistic influence exerted across densely embedded ties is greater than the linguistic influence exerted across other ties. In contrast, adding in F4 (local) only improved goodness of fit for three words: *asl*, *jawn*, and *lls*. Therefore, the support for hypothesis H2—that the linguistic influence exerted across geographically local ties is greater than the linguistic influence across than across other ties—is limited at best. Both the strength of ties and their locality are important factors for improvement in performance for words such as *jawn* and *asl*, which have a long history of being part of African American Vernacular English (AAVE) (Eisenstein 2017) and have been described as forming "street language" (Alim 2009). It is, therefore, no surprise that these words show improvement both due to the network and geographical cohesion in the way speakers organize.

3.7 Validity

The results empirically demonstrate that words diffuse in a social network as a contagion; furthermore, it is the strong ties, in comparison to local ties, that channel linguistic influence in the social network. These results, if taken on face value, suggest a fully endogenous mechanism at play in word diffusion. However, the models used in this work cannot differentiate the linguistic influence through social ties on Twitter from exogenous forms of influence, ranging from as simple as exposure to variation offline, exposure to speech on broadcasting media such as television, or exposure from other culturally important means such as music. While these are important distinctions that future work can tease out, the overall findings of this work still remain valid: irrespective of external influence, this work reveals influence pathways within the social network.

The limited number of geolocated users, and consequently the sparsity in the activation of the locality feature, in our dataset also raises another concern about the validity and generalization of the results. However, getting geolocation for only a small subset of Twitter users is quite common (e.g., Jurgens et al. 2015). A more subtle concern is that geolocation data can be biased towards certain demographics (e.g., Pavalanathan et al. 2015). Though well-founded, this concern is somewhat orthogonal to the substantive goal of this study, which focuses on finding the importance of shared geography, irrespective of multiple latent factors that could contribute to this shared geography.

3.8 Summary

This study demonstrates that language change in social media diffuses across a social network and is modulated by a number of sociolinguistic factors. Non-lexical words, such as phonetic spellings and abbreviations, spread in the form of a complex contagion: the likelihood of their adoption increases with the number of exposures. For both lexical and non-lexical words, densely embedded ties act as better conduits for transmission compared to other ties. In contrast, geographically local ties are only moderately effective. These findings indicate that language change is not merely a process of random diffusion over an undifferentiated social network, as proposed in many simulation studies (Fagyal et al. 2010; Griffiths et al. 2007; Niyogi et al. 1997). Rather, some social network connections matter more than others, and social judgments have a role to play in modulating language change. In turn, this conclusion provides large-scale quantitative support for earlier findings from ethnographic studies.

This study also demonstrates the use of computational models on diachronic text for the study of language change. The substantive findings of this study and the methodological modifications to the HP modeling framework that were necessary to test the hypotheses not just directly help answer RQ1 posed in § 1.1 but are also coherent with the theme set out with the broad substantive and methodological questions about linguistic leadership in language change in chapter 1.

Finally, HP models can be used for quantifying social influence in other social networks and for linguistic or non-linguistic applications. However, HP is a generative probabilistic model that makes certain parametric assumptions about how the cascade data is generated. The efficacy of HP models to detect social influence from event cascades is not known when the assumptions are violated. This can be potentially limiting to the adoption of HP models as a framework to study language change in social networks. The next chapter takes a general look at this problem and suggests an approach to overcome this limitation.

CHAPTER 4
SOCIAL INFLUENCE DETECTION IN EVENT CASCADES

4.1 Motivation

Peer influence or contagion is the reason for the spread of many online social behaviors (Crane 1999; Karsai et al. 2014; Yang et al. 2010). In chapter 3, I described how language change in online networks, pertaining to the adoption of non-standard words, can be viewed as a contagion on a network. But, despite its ubiquity, detecting peer influence from a network cascade is a challenging — or, in some cases, nearly impossible — enterprise in causal inference using just observational data (Shalizi et al. 2011). This is because network correlated behaviors, an outcome of peer influence, can also arise in its absence due to factors such as homophily— the tendency of individuals in a network to link with each other because of their similarity (McPherson et al. 2001)—and exogenous events because of sharing the same environment. To overcome these challenges, a slate of methods have proposed approximating the observational setting to a randomized experiment with varied success (Anagnostopoulos et al. 2008; Aral et al. 2009; La Fond et al. 2010).

However, a key question is whether existing methods are robust to phenomena such as missing data and misspecification. In realistic scenarios ranging from social media analysis to criminology, researchers lack complete records of event cascades: social media platforms may provide only a small sample, or relevant data may go unreported. Second, parametric models such as the Hawkes process require specifying hyperparameters, such as the timescale of the cascade. In real scenarios, there may be only a single cascade, making cross-validation impossible. These concerns motivate the following central research question for this work.

Can we develop a computational approach to detect peer influence in event cascades on a network to both account for the presence of a confound such as homophily, and be robust to practically important considerations of missing data and misspecification?

This chapter presents a simple Granger-style computational modeling test for social influence, which builds on *predictive* analysis of event cascades to answer the above research question.

Citation Sandeep Soni, Shawn Ling Ramirez, and Jacob Eisenstein. "Detecting social influence in event Cascades by comparing discriminative rankers". In: *The 2019 ACM SIGKDD Workshop on Causal Discovery*. 2019, pp. 78–99

4.2 Method

The setup is similar to chapter 3, meaning that an event cascade and a network are the only observational data. Each individual event in the cascade is a pair of timestamp and source, indicating when the event happened and which node in the network is associated with it. The core of the proposed approach is an online, discriminatively trained ranker, which learns to assign a high rank to nodes that are likely to host the next event in a cascade. The scoring function that the ranker uses is feature-based and incorporates both node and dyadic features. The score for node i at time t is given by,

$$\Psi_i(t; \boldsymbol{\theta}; \boldsymbol{\phi}) = h(\boldsymbol{g}_i; \boldsymbol{\phi}) + \sum_{e:t_e<t} \boldsymbol{\theta} \cdot \mathbf{f}_{s_e \to i} \kappa(t - t_e; \omega), \tag{4.1}$$

where both $\boldsymbol{g}_i$ and $\mathbf{f}_{\cdot \to i}$ are node and dyadic feature vectors for node i, and e indicates an event with source s_e and time t_e.

Node and dyad features The features $\boldsymbol{g}$ capture node level properties as vectors such as node embeddings (Grover et al. 2016; Tang et al. 2015), which are further transformed through a neural network h. The transformed features are intended to act as a proxy for homophily, which might otherwise confound the detection of social influence. Even observed covariates related to homophily (e.g. political party of a legislator, or the age of members in a social network), when available, can be incorporated in $\boldsymbol{g}$. The features $\boldsymbol{f}$, on the other hand, summarize dyad-level interactions a node is part of, such as self-activation or exposure to past participants which proxies social influence. The dyad features are multiplied by a decay kernel κ, and aggregated over the entire history of past events. The decay kernel captures the intuition that events in the distant past should have less impact on the score than more recent events.

Parameter estimation The scoring function has similarities to the intensity function of Hawkes process in § 3.4 with two key differences: the score can be negative and complicated integrals to estimate the survival probability are not needed to be calculated. The parameters of the scoring function are learned by minimizing the pairwise loss called WARP (Weighted Approximate Rank Pairwise; Usunier et al. 2009) between the predicted host of the ranker and the true host at the time of every new event. This enables the ranker to make online updates to the parameters and make predictions at every time, which is an added advantage. To avoid evaluating a quadratic number of node comparisons, the well-known WSABIE approximation (Weston et al. 2011) is used, which samples pairs until a violation is found.

Model comparison The operational methodology for testing peer influence involves training two competing rankers on a cascade—a baseline ranker and an experimental ranker—that differ only in the features used to make predictions. Specifically, the baseline ranker has all the features except the network exposure feature *i.e.* a dyadic feature indicating a tie with previous host of an event in the cascade. In contrast, the network exposure feature is additionally accounted by the experimental ranker. Each ranker is then applied to heldout data, and evaluated according to a ranking metric. To determine whether the performance difference between the two rankers is unlikely to have arisen due to chance alone, a non-parametric permutation test is applied for checking statistical significance (Smucker et al. 2007). The test repeatedly exchanges (permutes) the predictions between the two rankers to create an empirical distribution of the difference in ranker performance under the null hypothesis that the two rankers are identical. This equivalently tests the null hypothesis that adding social influence features does not improve ranking accuracy on heldout data, which is the Granger-based proxy of the null hypothesis of no social influence.

The combination of the discriminative ranker and the permutation test for social influence has two main advantages over contemporary models. First, the ranker makes no assumption about the data generation process which is a huge limitation of similar parametric models. This helps avoid potential issues arising from missing data and misspecification. Second, the ranker is light-weight to train but can still incorporate highly expressive features like node embeddings, which enables accounting for homophily. The next section describes the setup used to evaluate the ranker on the task of detecting social influence.

4.3 Evaluation

To determine the validity and efficacy of the proposed test for social influence, we evaluate it on a set of synthetic cascades generated over a real social network. Some cascades are generated without social influence, to test the Type I error rate of our method (incorrect rejection of the null hypothesis); other cascades are generated with social influence, to test the power (correct rejection of the null hypothesis). In both cases, the impact of homophily, self-excitation, missing data, and model misspecification is considered. The choice of synthetic data makes it possible to quantify these characteristics precisely.

Data generation Event cascades are generated using a multivariate Hawkes Process (HP) (see Equation 3.3) with the parameters tuned specifically to simulate conditions of social influence and various confounds. More concretely, the parameters are specified as follows:

$$\alpha_{j \to i} = a\mathbb{1}[j = i] + bA_{j \to i}\mathbb{1}[j \neq i] \tag{4.2}$$

$$\mu_i = \sigma\left(\beta v_i^{(2)} + \eta\right) \tag{4.3}$$

where $v^{(2)}$ is the second eigenvector of the Laplacian matrix of the network; $\sigma(\cdot)$ is the sigmoid function, which ensures that the base rate is positive; and A is the adjacency matrix of the network. Cascades can then be generated under various conditions of interest as:

`Social influence.` Setting $b = 0$ generates cascades without social influence, which are used to measure the Type I error rate of our test. As b increases, so does the impact of social influence. Throughout the evaluation, b is varied between 0 to 1 in increments of $.1$, depending on the experimental setting.

`Homophily.` To make similar nodes adopt at a similar rate, the base adoption rate of nodes is conditioned on their spectral components. This facilitates the generation of cascades with strong network correlation, even without social influence, corresponding to the case in which nodes form ties due to shared interest, and then participate in cascades that reflect that same interest. By varying the parameter β, the importance of homophily in shaping the cascades can be increased or decreased. β is varied between 0 to 7 in increments of 1, depending on the experiment condition. In all experiment settings, η is set to -5 for all the nodes.

`Self excitation.` In some cascades, nodes can participate repeatedly. This tendency can be controlled by setting $a > 0$. a is varied between 0 to 1 in increments of $.25$, depending on the experiment condition.

`Network structure.` Rather than generating a synthetic network, a real "mention" network from Twitter is considered. This static and directed network was constructed by first selecting all individuals who used a partisan political hashtag (e.g., `#clintonkaine2016`) between October 1-15, 2016, followed by identifying all individuals who were mentioned by someone in the initial set. The directed edge $i \to j$ indicates that i mentioned j on Twitter during this time period. The largest weakly connected component is selected as the underlying network for all synthetic cascades.

Evaluation metrics The ranking test is evaluated on two metrics: *validity* and *power*. A test is statistically valid if its p-values are well calibrated: at the threshold $p = \alpha$, the test should reject a true null hypothesis with probability less than or equal to α. To establish validity, the test is evaluated on cascades where the null hypothesis is known to be true. The rate at which the test rejects the null hypothesis is the Type I error rate. On the other hand, a test has high statistical power if it consistently rejects a false null hypothesis. Failure to reject a false null hypothesis leads to a Type II error and the power is the probability of not making a Type II error. To establish power, the test is evaluated on multiple cascades where the null hypothesis is known to be false. The rate at which the test fails to reject the null hypothesis is the Type II error rate. An ideal test should be valid under all conditions and should have high statistical power.

Baselines The performance of the ranking test is compared against two alternatives: (1) the shuffle test from (Anagnostopoulos et al. 2008), and (2) a test that compares the goodness of fit between two parametric Hawkes Processes: one with access to social influence parameters, and one in which these parameters are clamped to zero (see § 3.4 for details). These are referred in discussing the results as shuffle and HP respectively, whereas the discriminative rank comparison test is referred as ranking.

4.4 Results

Full data First consider the case when every test has access to full data—all the events in the cascade and all the edges in the network are known to each test—and the generative process is correctly specified, meaning that the exponential decay kernel and the bandwidth

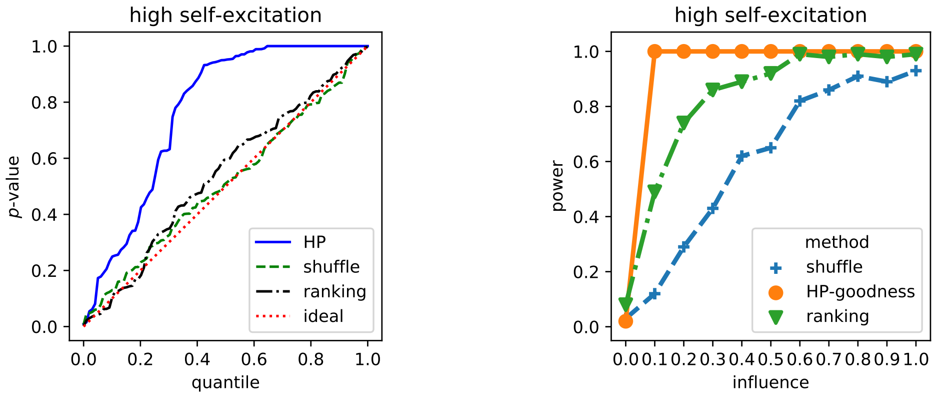


Figure 4.1: **Calibration plots.** (Left) Quantile-Quantile plots of p-values for all the three tests under no social influence, high homophily, and high self-excitation. The dotted red line shows the expected plot for an ideal well-calibrated test. (Right) Power for all the three tests under positive social influence, high homophily and high self-excitation.

parameter that modulates the rate of generation of future events are known during learning and testing. To check for validity, 100 cascades of 5000 events each are generated under conditions of no social influence ($b = 0$), varying homophily ($\beta > 0$) and varying self-excitation ($a > 0$). The null hypothesis of no social influence is true by design for every such cascade. Figure 4.1(left) shows the calibration of the tests when there is high homophily ($\beta = 7$) and high self-excitation ($a = 1$). Both ranking test and shuffle test are well-calibrated; HP test produces conservative p-values, but satisfies the condition of validity, which is that the Type I error rate is bounded by p-value. Conversely, to check for statistical power, 100 cascades of 5000 events each are generated by varying social influence at fixed values of homophily and self-excitation. As shown in Figure 4.1(right), the power increases with social influence for all tests, as expected. The HP test is the most powerful across these conditions, and the shuffle test is least powerful. Since the HP captures the true generative process, it should outperform the shuffle test, which is agnostic to the generative process. Because the ranking test also outperforms the shuffle test, in the subsequent evaluations it is only compared to HP.

Model misspecification Next how the tests fare under misspecification is examined, focusing on the bandwidth parameter (ω) for the temporal decay kernel. Again, 100 cascades

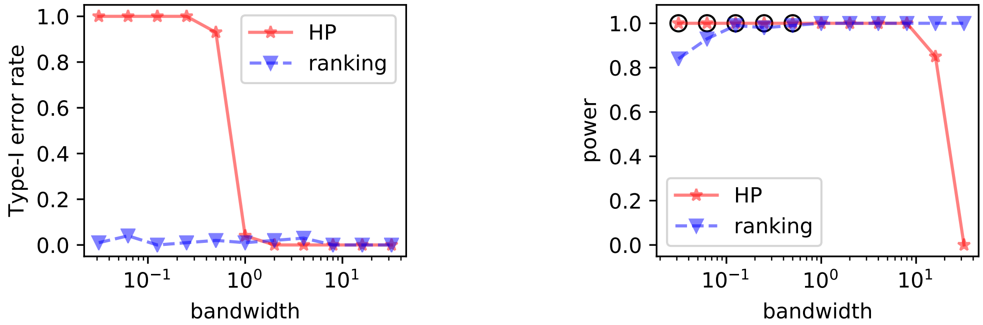


Figure 4.2: **Robustness against misspecification.** Type I error rate (left) and power (right) for HP and ranking test under misspecification of kernel bandwidth(ω). The HP test is invalid when $\omega < 1$; this is indicated by black circles around relevant data points when showing power.

of 5000 events each were generated for the influence ($b = 1$) and no-influence condition ($b = 0$), both with highest homophily and self-excitation ($a = 1$; $\beta = 7$). During generation, the bandwidth was fixed ($\omega^* = 1$), but both the HP learner and the ranker are unaware of this true value of the bandwidth parameter. The values of ω used by these tests was tuned to understand their effect on validity and the statistical power of both the tests. Varying the bandwidth parameter has a natural interpretation: as ω decreases, the scope of the history is effectively widened; increasing this parameter has the opposite effect. The results for both validity and power are shown in Figure 4.2.

The HP model is quite sensitive to this parameter, and misspecification severely undermines the validity of the test for practical purposes. A high bandwidth value keeps the Type I error rate in check but limits power as the effect of past exposure history is nullified. But a low bandwidth parameter makes irrelevant neighbor activations deemed consequential affecting Type I error rate in no-influence scenarios. In contrast, the ranking test is robust to misspecification: for a wide range of bandwidth values, the ranker maintains a low Type I error rate and has considerable power. This is because the ranking objective requires only that the relative order of each node is maintained. Spurious events in the distant past affect many nodes, and the resulting changes in rank are not significant. While there is research on augmenting the HP with nonparametric kernels that are learned from data (e.g., Zhou

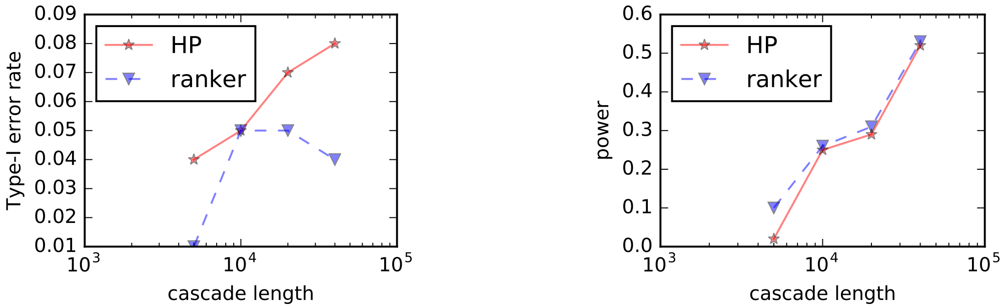


Figure 4.3: **Robustness against missing events.** Type I error rate and power for HP and ranker test on 99 % random missing events on cascades of increasing length. The ranker's power is always higher than HP particularly on cascades of shorter length (5000 events). The Type I error rate for both methods is around 0.05.

et al. 2013), such methods are complex to implement, and require large amounts of training data. For a parametric kernel, the bandwidth can be learned by cross-validation, but only if multiple cascades are available and if the parameter is guaranteed to be static over time. In contrast, the ranking test can easily incorporate multiple kernels into the discriminative ranking function and is completely online, thereby having the flexibility to learn complex triggering patterns from a single cascade.

Missing data In many real situations, the assumption that all events in a cascade are available is easily violated. For example, high expense in collecting data, rate limits in data collection such as Twitter's public API, changes to data privacy settings, data loss, and late start to data collection are all but a few examples. Incomplete data can diminish statistical power, giving the appearance of unprompted innovations to events that were in fact socially motivated. To quantify this phenomenon, cascades with two different types of missing events were generated: events missing at random, and events missing in contiguous blocks.

For random missing events, 100 cascades of varying lengths were generated, and events were dropped randomly at a specified rate of 99% before presenting them to the learners. This sampling rate matches a published estimate of the fraction of tweets included in Twitter's streaming API Morstatter et al. (2013) .Figure 4.3 shows the power of the HP and ranking tests, as a function of cascade length. While both tests increase in power with the cascade length, the HP is marginally less powerful across all cascades, particularly on short cascades. The results are shown in Figure 4.3.

Xu et al. (2017) consider an alternative scenario, in which contiguous blocks of events are missing both at the start and the end of the cascade, resulting in *short doubly censored* sequences, which is relevant to social media analysis due to API limits (Xu et al. 2017). To simulate this, 100 cascades of different lengths were generated and 99% of events were dropped, equally split between the two ends of a cascade and in a contiguous block. On censored cascades, the HP test suffers from a high Type I error rate, limiting its validity. This bias towards overestimating social influence is consistent with findings from Xu et al. (2017) , who point it is due to overfitting. Conversely, the validity of the ranking test was not affected by censoring, because the test does not rely on explaining the temporal distribution of events. However, the loss of information due to censoring does affect the power of the ranking test, especially if the length of the cascade is small; this is a natural consequence of the difficulty of estimation from limited data.

In summary, the evaluation on synthetic data shows that the shuffle test is underpowered, because it uses only the first-time activations for each node, and because it ignores the time between activations. The HP test has low Type I error rate and high power when (a) the model is correctly specified and (b) complete data is available. However, it is highly sensitive to misspecification of the time kernel, and to missing events. These cases can make the test statistically invalid or hurt its power. The ranking test is robust across all scenarios: it is valid in all scenarios, and nearly matches the HP test's power when complete data is available; model misspecification has little impact on its validity or power; and it is reasonably powerful even under different types of missing data conditions.

4.5 Real world applications

To assess the effectiveness of the ranking test in the real world, two case studies where influence is suspected are considered: first, a non-linguistic setting of cosponsorship decisions by legislators viewed as cascades on a political finance network; second, a linguistic setting of rumors on Twitter pertaining to the discovery of the Higgs Boson particle.

Cosponsorship and donor influence A key step in the legislative process is when a bill receives endorsement from legislators besides its original author, called cosponsors. Cosponsorship decisions are important markers of wider support; they signal the expertise of the original sponsor, and provide information about the bill's content and the party, ideological, or constituency base for whom the bill advocates. However, cosponsorship is also a low-cost means of position taking, reflective of favor-trading, vote-buying, and special interest politics (Kessler et al. 1996). An open question is whether cosponsorship decisions are influenced by campaign donations, for example by facilitating special access to legislators (Kalla et al. 2016). To test this question, an affiliation network is constructed among legislators if they share common campaign donors, and the discriminative ranking test is applied to sequences of cosponsorship decisions. The detection of influence on this network would be compatible with the hypothesis that campaign donations influence cosponsorship decisions. Using cosponsorship sequences for bills introduced in 115th Congress viewed as a separate cascade, and a network of legislators constructed by linking them if they share a campaign donor, evidence shows that cosponsorship decisions can be predicted more accurately by considering exposures over the donor network in addition to party and state affiliation which proxy shared interests. The result is shown in Figure 4.4.

Rumor cascades The ranking test is also applied to determine whether social influence and homophily were factors in the spread of rumors related to the Higgs Boson discovery, an interesting case study since the rumors were triggered by an exogenous event. In the

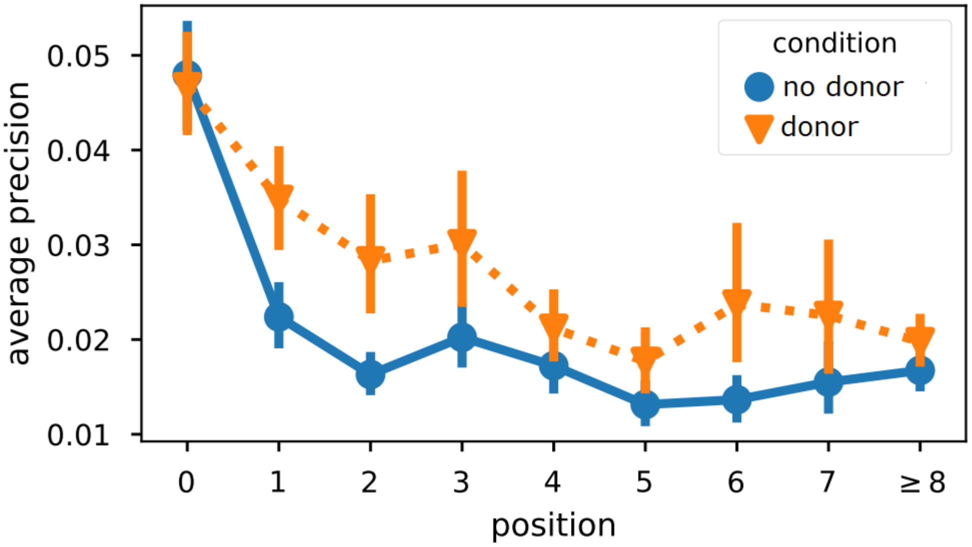


Figure 4.4: **Donor influence in bill co-sponsorship.** Ranking performance for the legislation co-sponsorship cascades, organized by the position of each co-sponsor in the cascade. Position 0 is the initial sponsor.

absence of ground truth, the quality of predictions are compared with chance in the first baseline as users are ranked randomly. In the second baseline, users are ranked by their number of past events (tweets about the Higgs boson); users with no such events are ranked by network degree, with the intuition that this is a rough proxy for their rate of tweeting. The predictions improve over both the baselines indicating that both homophily and social influence could be responsible in making these rumors cascade.

4.6 Summary

Results from experiments on both synthetic and real data suggest that the discriminative ranking approach to detect social influence in event cascades has many merits over existing methods. It is also an explanatory model with predictive ability, a characteristic argued to be more desirable for social scientific methods (Hofman et al. 2017). However there are also a few limitations. First, Granger causality, the foundational basis for the model, is a limited, *predictive* notion of causality; Granger causality in an event cascade does not necessarily imply that an intervention into the cascade at time t would affect subsequent events. Next, the approach hinges on the assumption that homophily is effectively proxied. When such an assumption fails, it is generally not possible to distinguish social influence from homophily using observational data (Shalizi et al. 2011). Note, however, that the discriminative ranker model has some capacity to handle unobserved confounds through its use of node embeddings in a manner similar to (Veitch et al. 2019). Finally, the ranker comparison is based on a permutation test, which is a non-parametric distribution-free approach. When the predictions of the two rankers are very similar, the resulting distribution can have low variance, leading to conservative p-values. This problem is especially severe in three cases: when the network size is small; when the heldout set is small; and when the base ranker is already very accurate, leaving little room for improvement with the addition of the social features. The conservativeness of the permutation test does not affect the validity of the test (Edgington et al. 2007), but it can affect the power.

The method proposed in this part does not make any assumption about the content of the cascade *i.e.* it only requires a sequence of events and does not care what those events represent. The method can, therefore, be applied to linguistic data such as the one in chapter 3.

CHAPTER 5
SEMANTIC LEADERSHIP OF DOCUMENTS

5.1 Motivation

Elements of a natural language comprise of signs, such as words, and signifieds, the meaning associated to these signs (De Saussure 1916). Word embeddings are popularly used to represent the meaning of a word by leveraging the distributional hypothesis (Harris 1954; Firth 1957). However, as language evolves, new meanings get attached to existing words (Traugott et al. 2001). Researchers have suitably extended word embeddings to *diachronic* word embeddings—word embeddings over time—using methods ranging from statistical techniques (Kulkarni et al. 2015; Hamilton et al. 2016b; Yao et al. 2018) and Bayesian approaches (Wijaya et al. 2011; Frermann et al. 2016; Bamler et al. 2017; Rudolph et al. 2018), to using neural networks (Kim et al. 2014; Rosenfeld et al. 2018). Diachronic word embeddings can identify semantic changes by quantifying and measuring the shifts in each word's distributional neighborhood (Kutuzov et al. 2018). But, even though these models are effective in identifying words that have changed, they treat each word in isolation. On their own, they lack the ability to indicate *where* the change takes place and *who* is at the forefront of the change: which documents or passages are at the leading edge of semantic change, and which lag behind?

The ability to zoom into a document and find the extracts of semantically progressive language use in the document can be useful in domains such as digital humanities by providing more granularity in distant reading settings (Jänicke et al. 2015). Alternatively, identifying precisely the leading and lagging documents based on the semantics is not just a methodological challenge in itself but also the first step in assessing the impact of these documents. As an example, this capability would provide new evidence about the social processes underlying linguistic and scholarly influence (Gerow et al. 2018). Such evidence could be used to examine any social inequities in Science.

The main research questions for this work are:

Q1 In a large timestamped collection of documents, can computation methods identify instances of semantic changes that are used progressively *i.e.* with an advanced meaning at the time?

Q2 If a document contains multiple progressive instances can the document be scored for its progressiveness by aggregating over individual instances?

Q3 Can semantically progressive documents be tested to assess if they are influential?

Citation Sandeep Soni, Kristina Lerman, and Jacob Eisenstein. "Follow the leader: Documents on the leading edge of semantic change get more citations". In: *Journal of the Association for Information Science and Technology* 72.4 (2021), pp. 478–492

5.2 Method

Diachronic word embeddings make it possible to measure lexical semantic change over time (e.g., Kulkarni et al. 2015; Hamilton et al. 2016b). In standard word embeddings, each word type is associated with a vector of real numbers, based on its distributional statistics (Turney et al. 2010; Mikolov et al. 2013a). In diachronic word embeddings, this vector is time-dependent, reflecting how a word's meaning (and associated distributional statistics) can change over time. Building on diachronic word embeddings, the method in this chapter is comprised of four steps: (1) learning diachronic embeddings of words; (2) identifying semantic innovations using their diachronic embeddings; (3) scoring each usage by its position with respect to the semantic change; and (4) aggregating these scores by document. A schematic of the entire pipeline is shown in Figure 5.1. A detailed description of each of these steps follows.

5.2.1 Estimating word embeddings

Formally, assume a finite vocabulary $\mathcal{V}$, and two corpora, $\mathcal{W}^{\text{(old)}}$ and $\mathcal{W}^{\text{(new)}}$, where each corpus is a set of sequences of tokens, $\mathcal{W} = \{(w_{i,1}, w_{i,2}, \ldots, w_{i,T_i})\}_{i=1}^{N}$, where N is the number of documents in the corpus, i indexes an individual document whose length is T_i, and each $w_{i,t} \in \mathcal{V}$. For each corpus, a set of word embeddings on the single vocabulary $\mathcal{V}$ is estimated. Following Hamilton et al. (2016) , skipgram embeddings (Mikolov et al. 2013a) are estimated, which are based on the objective of predicting context words $w_{t'}$ conditioned on a target word w_t. Omitting the document index i, the skipgram objective is based on the probability,

$$P(w_{t'} \mid w_t) \propto \exp\left(\mathbf{v}_{w_{t'}} \cdot \mathbf{u}_{w_t}\right), \tag{5.1}$$

where $\mathbf{v}_{w_{t'}}$ is the embedding of $w_{t'}$ when it is used as a context (also called as the "output" embedding), and $\mathbf{u}_{w_t}$ is the embedding of w_t when it is used as a target word (also called as the "input" embedding). Normalizing the probability in Equation 6.1 requires summing over all possible $w_{t'}$, which is computationally expensive. Negative sampling is one solution (Mikolov et al. 2013a), but this does not yield properly normalized probabilities. Noise-contrastive estimation (NCE; Gutmann et al. 2010), which makes it possible to estimate the probability in Equation 6.1 without computing the normalization term (Mnih et al. 2013), is used instead. The input embeddings between the corpora $\mathcal{W}^{\text{(old)}}$ and $\mathcal{W}^{\text{(new)}}$ are reconciled to make them comparable across the two corpora following the step proposed by Hamilton et al. (2016) , which applies the Procrustes method (Gower et al. 2004) to identify an orthogonal projection $\mathbf{Q}$ that minimizes the Frobenius norm $||\mathbf{Q}\mathbf{U}^{\text{(old)}} - \mathbf{U}^{\text{(new)}}||_F$, where $||\mathbf{X}||_F = \sqrt{\sum_{i,j} x_{i,j}^2}$.

Sensitivity to initialization One potential downside of NCE is that its embeddings depend on the random initialization, unlike deterministic techniques such as singular value decomposition (Sagi et al. 2011; Levy et al. 2014). As a result, the list of near neighbors can change across multiple runs (Hellrich et al. 2016). Nonetheless, NCE is chosen because the resulting embeddings outperformed alternatives on intrinsic word similarity benchmarks (Luong et al. 2013). Multiple robustness checks indicated that the method identified similar sets of semantic innovations across multiple runs.

5.2.2 Discovering semantic innovations

After estimating the diachronic embeddings for each word, the next step is to identify semantic innovations: words that have shifted in meaning. One possibility would be to directly measure differences between the embeddings $\mathbf{u}^{(\text{old})}$ and $\mathbf{u}^{(\text{new})}$, but this can be unreliable because the density of embedding space is not guaranteed to be uniform. Instead, the local second-order approach proposed by Hamilton et al. (2016) is followed. In their protocol, first, a union of the sets of a word's near-neighbors ($n = 50$) in the "old" and "new" periods is taken. Next, the similarity of the word's embedding to the embeddings for members of this set is calculated, for both the "old" and "new" embeddings. This yields a pair of vectors of similarities, each reflecting the word's position in a local neighborhood. The degree of change in a word's position is the distance between these two vectors.

5.2.3 Situating progressive usages of semantic innovations

Given a set of semantic innovations $\mathcal{S} \subset \mathcal{V}$, the methodological innovation of this work is to situate usage with respect to semantic changes. Each usage of an innovation $w^* \in \mathcal{S}$ can be analyzed using the likelihood function underlying the skipgram objective, and scored by the ratio of the log-likelihoods under the embedding models associated with $\mathcal{W}^{(\text{old})}$ and $\mathcal{W}^{(\text{new})}$. Specifically, this turns out as

$$r_{w^*,i} = \sum_{t:w_{i,t}=w^*} \sum_{\substack{j \geq -k \\ j \leq k \\ j \neq 0}} \log \frac{P^{(\text{new})}(w_{i,t+j} \mid w^*)}{P^{(\text{old})}(w_{i,t+j} \mid w^*)}. \tag{5.2}$$

The intuition behind the statistic is to predict the context of every appearance of the semantic innovation w^* in the document i using both the "new" and "old" meaning of w^* and the surrounding context. These new and old meanings are obtained from the embedding models associated with $\mathcal{W}^{(\text{new})}$ and $\mathcal{W}^{(\text{old})}$ respectively. Substituting the form of probability from Equation 6.1 and simplifying further, the log-likelihood ratio reduces to:

$$\begin{aligned} r_{w^*,i} = \sum_{t:w_{i,t}=w^*} \sum_{\substack{j \geq -k \\ j \leq k \\ j \neq 0}} & \mathbf{v}^{(\text{new})}_{w_{i,t+j}} \cdot \mathbf{u}^{(\text{new})}_{w^*} - Z^{(\text{new})}_{w^*} \\ & - \mathbf{v}^{(\text{old})}_{w_{i,t+j}} \cdot \mathbf{u}^{(\text{old})}_{w^*} + Z^{(\text{old})}_{w^*}, \end{aligned} \tag{5.3}$$

where Z_{w^*} is the log normalization term, $\log \sum_{w'} \exp\left(\mathbf{v}_{w'} \cdot \mathbf{u}_{w^*}\right)$. This metric intuitively favors documents that use w^* in contexts that align with the new embeddings $\mathbf{u}^{(\text{new})}_{w^*}$ and $\mathbf{V}^{(\text{new})}$. This step is the proposed solution for `Q1`.

5.2.4 Aggregating to document scores

Given a set of innovations $\mathcal{S} \subset \mathcal{V}$, for each document i gets a set of scores $\{r_{i,w^*} : w^* \in \mathcal{S}\}$. The score for document i is the maximum over the set of innovations, $m_i = \max_{w^* \in \mathcal{S}} r_{i,w^*}$. This quantifies the maximal extent to which the document's lexical semantics match that of the more contemporary embedding model, $(U^{(\text{new})}, V^{(\text{new})})$. We then standardize against other documents published in the same year, by computing the z-score, $z_i = \frac{m_i - \mu}{\sigma}$, where μ is the mean score for documents published in the same year, and σ is the standard deviation. Documents with a positive z-score have lexical semantics that better match the contemporary embedding model than other documents written at the same time, and can thus be said to be semantically progressive. By standardizing each year separately, the progressiveness metric does not inherently favor older or newer texts. This proposed approach is a solution to `Q2`.

As a robustness check, we also experimented with an alternative discretized approach for scoring the documents. In this scheme, the document score was calculated as the num-

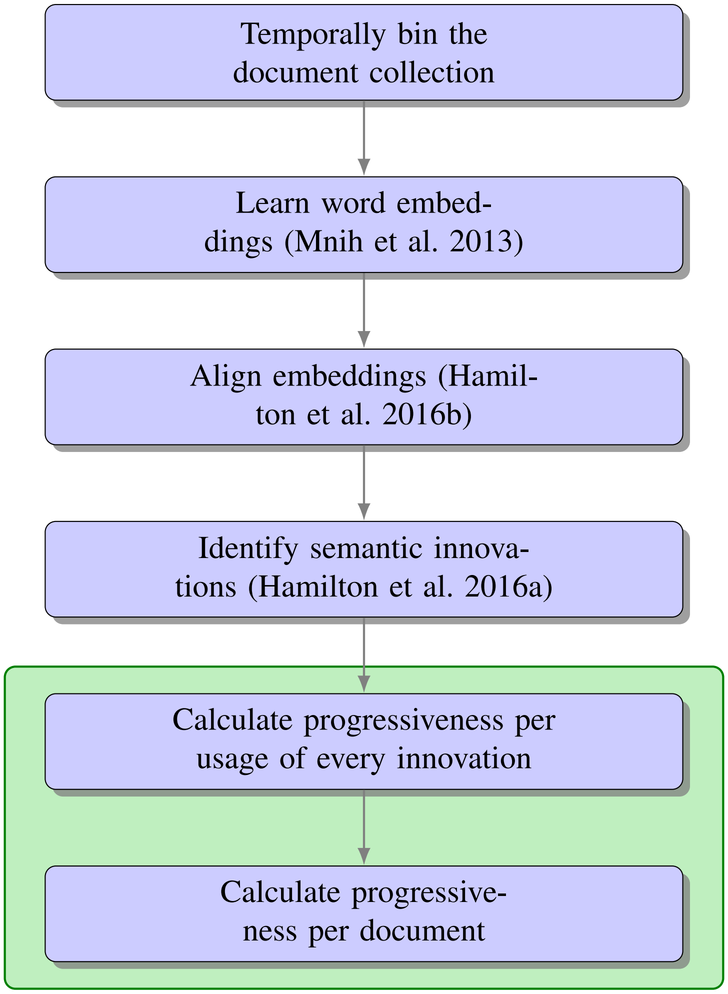

Temporally bin the document collection
Learn word embeddings (Mnih et al. 2013)
Align embeddings (Hamilton et al. 2016b)
Identify semantic innovations (Hamilton et al. 2016a)
Calculate progressiveness per usage of every innovation
Calculate progressiveness per document

Figure 5.1: **Method pipeline.** Flowchart shows our complete pipeline and highlights (in green) our methodological contributions.

ber of innovations whose progressiveness exceeds the median progressiveness value over the entire set of innovations. The subsequent analysis of the documents with this scoring scheme produced qualitatively similar results to those obtained with the measure described in the previous paragraph (innovativeness of maximally innovative word per document), and so are only included in the appendix B.

5.3 Data

For empirical validation of our approach, two document collections are considered: documents representing *legal opinions* in federal courts of the United States of America (Ler-

Table 5.1: **Descriptive statistics.** A summary of the two datasets that are considered in this study.

Statistic	Legal opinions	Scientific abstracts
Number of documents	3,854,738	2,408,010
Years	1754–2018	1949–2018
Average number of citations (in-degree)	7.84	39.19
Average number of references (out-degree)	7.80	9.49
Length (number of unique word types per document)	632.22	93.10

man et al. 2017),[1] and the DBLP collection of *computer science abstracts* (Ley 2002).[2] These datasets were chosen because they include timestamps as well as citation information, making it possible to link semantic innovation with influence in a citation network for large-scale quantitative evaluation.

Legal opinions. A legal opinion is a document written by a judge or a judicial panel that summarizes their decision and all relevant facts about a court case. All legal opinions are obtained by using the bulk API of a publicly available service.[3] These opinions span over 400 courts, multiple centuries and have a broad jurisdictional coverage.

Scientific abstracts. The abstracts from DBLP were obtained from ArnetMiner,[4] a service that has released multiple versions of this data with the latest papers since 2010 (Tang et al. 2008; Sinha et al. 2015). The latest version (v10) from their collection was chosen.

Metadata. Both datasets feature common metadata, including the year in which the document was published, the number of citations the document has received and the number of references to other documents in the citation network. A descriptive summary of the complete collection is given in Table 5.1.

[1] https://www.courtlistener.com/

[2] https://dblp.uni-trier.de/

[3] https://www.courtlistener.com/api/bulk-info/

[4] https://aminer.org

5.4 Results

5.4.1 Qualitative results

Semantic changes

A few prominent semantic innovations are listed in Table 5.2. The innovations in the legal opinions corpus we discover span multiple domains, including financial (e.g. *laundering*, which earlier exclusively meant washing), socio-political (e.g. *noncitizens*, which was earlier close to *tribals* or *indians* but has now moved closer in meaning to *immigrants*), medical (e.g. *fertilization*, which was first used in the context of agriculture, but now increasingly refers to human reproduction) and technological (e.g. *web*, which now refers almost exclusively to the internet). The analysis also independently discovers semantic changes in words like *cellular* and *asylum*, which have previously been identified as semantic changes in other corpora (Kulkarni et al. 2015; Hamilton et al. 2016b; Hamilton et al. 2016a). In the scientific domain, a common source of semantic innovation is through the use of abbreviations. Examples include *nfc*, which earlier meant "neuro-fuzzy controllers" but lately refers to "near-field communication"; *ux*, which was used as a short form for unix, but is now increasingly used to mean "user experience"; and *ssd*, which popularly stood for "sum-of-squared difference", but of late additionally means "solid state drives." Another common source of semantic innovations is the creative naming of technological components. Examples include *cloud*, which now refers to services offered through the internet in comparison to its mainstream meaning; *spark*, which was earlier popularly used to mean ignition, but has lately been referred to the popular MapReduce framework; and *android*, which referred to robots with human appearances, but now commonly refers to the mobile computing operating system.

Table 5.2: **Semantic innovations.** Examples of semantic innovations identified by the proposed method for both the datasets.

Doc. type	Innovations	Old usage	New usage	Example document with new usage
Legal	*laundering*	***laundering*** *clothing*	***laundering*** *funds*	United States v. Talmadge G. Rauhoff, (7th Cir. 1975)
	asylum	*insane* ***asylum***	*political* ***asylum***	Bertrand v. Sava, (S.D.N.Y. 1982)
	fertilization	*soil,* ***fertilization***	*post-****fertilization*** *contraceptive*	Planned Parenthood vs Casey (505 U.S. 833)
Science	*ux*	*hp-****ux***	*user experience* ***(ux)***	Hassenzahl et al. (2006)
	surf	***surf*** *the internet*	*descriptor* ***surf***	Bay et al. (2008)
	android	*intelligent* ***android***	*google's* ***android***	Shabtai et al. (2010)

Leading documents

Two examples of legal opinions at the leading edge of change according to the proposed metric are Planned Parenthood vs Casey (505 U.S. 833) and United States v. Talmadge G. Rauhoff, (7th Cir. 1975). The landmark 1992 opinion in Planned Parenthood vs Casey was identified by our method as leading a change with several semantically progressive terms like *fertilization*, *provider*, and *viability* mentioned in the document. The term *fertilization* had previously been used in the context of agriculture, but this decision was an early example of an increasingly common usage in connection with reproductive rights:

- *. . . two-week gestational increments from* ***fertilization*** *to full term . . .*
- *. . . before she uses a post-****fertilization*** *contraceptive.*

Similarly, the United States v. Talmadge G. Rauhoff, (7th Cir. 1975) scores highly on our measure and was one of the first to use *laundering* to refer to illegal transfer of money:

- *. . . $15,000 as part of the* ***'laundering'*** *process . . .*
- *. . . first step in the successful* ***laundering*** *of the funds. . .*

The first mention of the term was quoted, which may indicate a metaphorical intent.

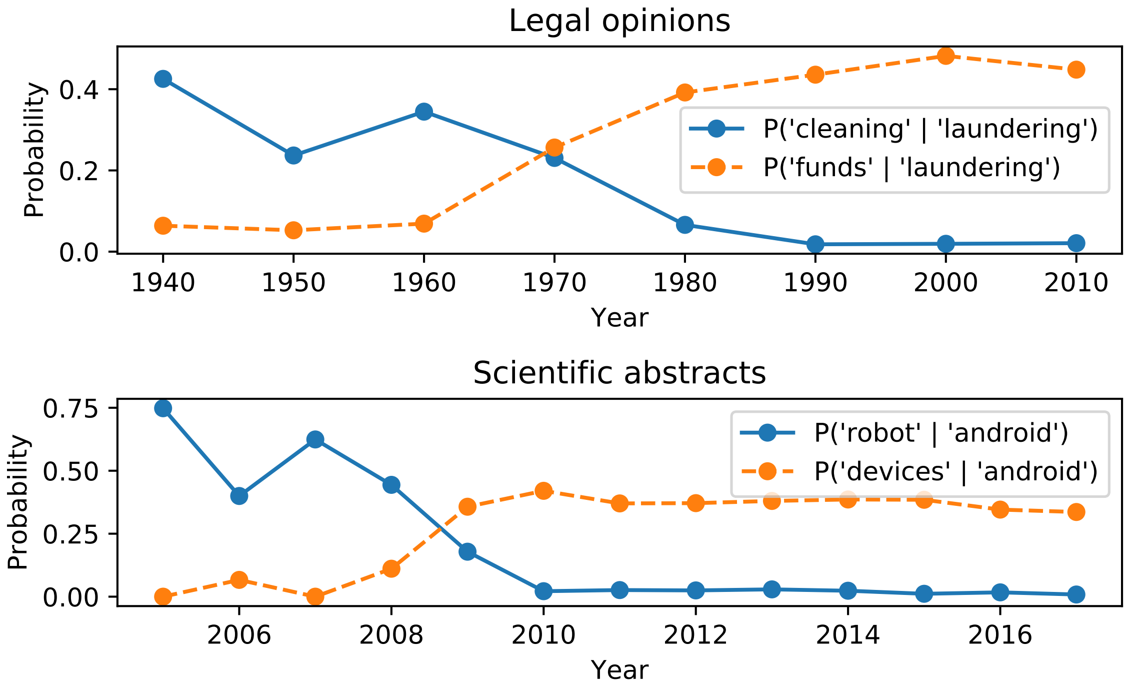


Figure 5.2: **Examples of semantic changes identified by the method.** In the upper time series, the meaning of the term *laundering* evolves to include money laundering, shown here as the increase in the conditional probability of seeing the term *laundering* given that *funds* appears in a document, in contrast to the conditional probability of *laundering* given that *cleaning* appears. For this change, a prominent leading document is the opinion in U.S. v Rauhoff (1975). In the lower time series, the meaning of *android* evolves to include the mobile phone operating system. A prominent leader of this change is (Shabtai et al. 2010).

In the scientific domain, the seminal paper on the Android operating system is rated as a semantically progressive document (Shabtai et al. 2010). At that time, the conventional meaning of the term *android* was an interactive robot (e.g. . . . *interaction using an* ***android*** *that has human-like appearance. . .*), but Shabtai et al. used the now-prevalent meaning as a mobile operating system (e.g. . . . *the* ***android*** *framework* . . .). Figure 5.2 shows the evolution of the semantic innovations which approximately aligns with the leading documents that our method discovered.

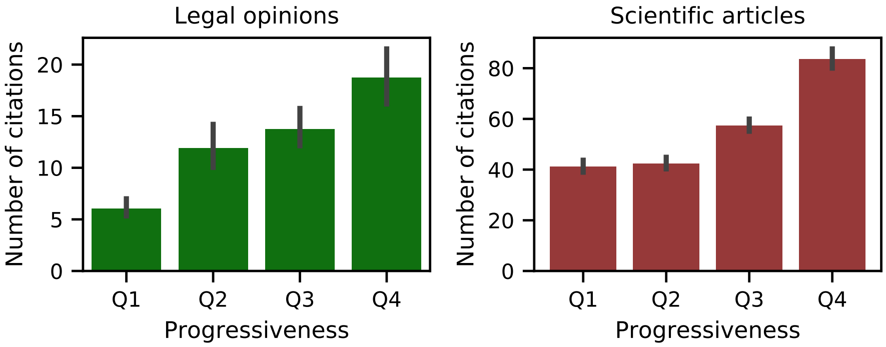


Figure 5.3: **Linking semantic progressiveness with citation impact.** The univariate relationship between the number of citations and our measure of semantic progressiveness. For both the legal opinions and the scientific articles, the citations increase for more progressive documents.

5.4.2 Quantitative evaluation

Lacking large-scale manual annotations for the semantic progressiveness of legal opinions or scientific abstracts, *influence* is measured, as quantified by citations. Specifically, we investigate the hypothesis that more citations will accrue to documents that our metrics judge to be semantically progressive.

Note that a one-way causal relationship from semantic innovation to influence is not hypothesized. Semantic progressiveness may cause some documents to be highly cited, due to the introduction or usage of helpful new terminology. But it also seems likely that documents that are well-cited for more fundamental reasons — e.g., significant methodological innovations in science, foundational precedents in law — will also exert an outsize effect on language. For example, the highly cited paper on Latent Dirichlet Allocation (Blei et al. 2003) introduced a new meaning for the term *LDA* (which also refers to linear discriminant analysis), but in this case it is likely that the underlying cause is the power of the method rather than the perspicacity of the name. The key point of these evaluations is to test the existence of a previously unknown correlation between language and citation networks, and to provide a further validation of our measure of semantic progressiveness.

Univariate relationship

Figure 5.3 shows the number of citations for each quartile of the progressiveness measure, indicating a steady increase in both datasets. This figure excludes documents that do not include any of the terms identified as having changing semantics. Also excluded are documents predating 1980, which skew the population with a few landmark examples with vast citation counts; these documents are included in the multivariate analysis that follows.

Multivariate relationship

Many factors drive citation counts, such as age, length, and content (Fowler et al. 2007; Van Opijneni 2012). Some of these factors may be correlated with semantic progressiveness, confounding the analysis: for example, older documents have more chances to be cited, but are unlikely to lead a semantic change that would be captured by our metrics. To control for these additional predictors, the problem can be formulated as a multivariate regression. The dependent variable is the number of citations, and the predictors include our measure of semantic progressiveness, as well as a set of controls. A Poisson regression model is fit since the number of citations is a count variable.[5]

Evaluation setup Two baseline models (M1 and M2) and two experimental models (M3 and M4) are constructed to systematically assess the relevance of semantic progressiveness to citation prediction.

M1 includes document covariates capturing structural information about each document: the number of outgoing references makes; its age; its length; number of authors (available only for scientific articles). In addition, another covariate from the prediction of a lightweight bag-of-words regression model on the document (similar to

[5] In cases of overdispersion (high variance), negative binomial regression is preferred to Poisson regression (Greene 2003). However, the Cameron-Trivedi test (Cameron et al. 1990) did not detect overdispersion in our data.

Table 5.3: **Poisson regression analysis of citations to scientific abstracts.** Each column indicates a model, each row indicates a predictor, and each cell contains the coefficient. Log likelihood is in millions of nats.

Predictors	M1	M2	M3	M4
Constant	1.983	1.943	2.032	1.770
Outdegree	0.009	0.009	0.009	0.009
# Authors	0.055	0.054	0.054	0.054
Age	0.079	0.079	0.078	0.073
Length	0.002	0.002	0.002	0.002
BoWs	0.000	0.000	0.000	0.000
# Innovs		0.028	-0.010	-0.034
Prog.			0.137	
Prog. Q2				0.179
Prog. Q3				0.431
Prog. Q4				0.698
Log Lik.	-13.07	-13.06	-12.93	**-12.82**

Yogatama et al. 2011) is included, with the intuition that some topics may get cited more than others.

M2 includes all the covariates from M1, and an additional covariate for the number of unique semantic innovations present in the document

M3 includes all covariates from M2, and an additional covariate for the semantic progressiveness score as the numerical value.

M4 includes all covariates from M2, and an additional 3 categorical covariates for the semantic progressiveness binned into quartiles.

The models are compared by goodness-of-fit, a standard technique from quantitative social science (Greene 2003), and the improvement in the goodness-of-fit for M3 and M4 over M1 and M2 is tested using a χ^2 test.

Results The regressions reveal a strong relationship between semantic progressiveness and citation count. For the scientific abstracts (Table 5.3), M3 and M4 obtain a significantly better fit than M1 ($\chi^2(2) = 137767, p \approx 0$ and $\chi^2(4) = 250479, p \approx 0$ respectively).

Table 5.4: **Poisson regression analysis of citations to legal documents.** Each column indicates a model, each row indicates a predictor, and each cell contains the coefficient.

Predictors	M1	M2	M3	M4
Constant	1.614	1.421	1.476	1.168
Outdegree	0.022	0.020	0.021	0.020
Age	0.009	0.011	0.010	0.010
Length	0.000	-0.000	-0.000	-0.000
BoWs	-0.000	-0.000	-0.000	-0.000
# Innovs		0.054	0.045	0.042
Prog.			0.094	
Prog. Q2				0.384
Prog. Q3				0.382
Prog. Q4				0.470
Log Lik.	-415195	-410601	**-406843**	-408031

M3 and M4 also obtain a significantly better fit than M2 ($\chi^2(1) = 130176, p \approx 0$ and $\chi^2(3) = 242889, p \approx 0$ respectively). The effect sizes are relatively large: the coefficient of 0.698 for top quartile of semantic progressiveness corresponds to an increase in the expected number of citations by a factor of 2, in comparison with documents in the bottom quartile.

The story is similar for the legal opinions (Table 5.4), with only minor differences. Both M3 and M4 significantly improve the goodness of fit over the baseline M1 ($\chi^2(2) = 8352, p \approx 0$ and $\chi^2(4) = 7164$ respectively) and the baseline M2 ($\chi^2(1) = 3758, p \approx 0$ and $\chi^2(3) = 2571$ respectively), indicating again that semantic progressiveness of the document is highly predictive of the number of incoming citations, even after controlling for several covariates. The coefficient of 0.47 for the top quartile of progressiveness corresponds to an increase in the expected number of citations by a factor of 1.6, as compared to the bottom quartile. Overall, these results indicate that the measure of semantic progressiveness adds substantial new information to the array of covariates included in the baseline models, and that semantically progressive documents receive significantly more citations.

5.5 Validity

There are a few issues related to validity that merit further clarification. First, the absence of gold labels to classify words as semantic changes — a common obstacle for research in semantic change detection — makes predictive evaluation difficult. However, recent research (e.g., Kaiser et al. 2020; Martinc et al. 2020) on a semeval task of semantic change detection (Schlechtweg et al. 2020) shows that the model for semantic change detection used in this study outperforms other competing models. Additionally, the robustness checks shown in appendix B add further credence to the obtained semantic changes in this study.

Second, the semantic change detection model cannot differentiate between drifts in word meaning to drifts in topics over time. For example, the model in this study detects that the term *race* shifts around the sixties in legal opinions towards a meaning that refers to human grouping and ancestry, despite the existence of this sense even a century ago. The shift, therefore, is likely due to topical changes during that time. Future work can try to tease these two apart but for the purpose of this study, which concerns with finding progressive usages of any word, change in meaning and change in content fall under the same purview.

Third, similar to semantic change detection, no ground truth is available to adjudicate if a document is semantically progressive or otherwise. Once again, this takes precision-recall style evaluation out of question. As an alternative, this study explores the relationship between being innovative or progressive in a linguistically as a proxy of linguistic influence and relates it to influence to a more objective and prevalent notion of influence in terms of the number of citations. Future work could relate linguistic influence to even more forms of influence. Alternative, future work could set up a cloze task with domain experts where the current model can suggest progressive documents that humans can evaluate.

Finally, as has been mentioned earlier the positive relationship between semantic progressiveness and the number of citations should not be misconstrued as a causal link between the two. Future work can attempt to further quantify the causal relationship between these two variables.

5.6 Summary

This work shows how to identify the leading examples of semantic change, by leveraging the models underlying diachronic word embeddings. This enables us to test the hypothesis that semantically progressive documents—that is, documents that use words in ways that reflect a change in progress—tend to receive more citations. This technique has potential applicability in the digital humanities, computational social science, and scientometrics (the study of science itself; see Van Raan 1997).

The only metadata considered in learning diachronic word embeddings in this work was the timestamps of the documents. In many situations, additional metadata may be available, such as the identity of the producer of documents, the community the documents belong to, the location the documents are from, *etc*. Models that can take take this additional metadata and specialize the diachronic word embeddings are advantageous because they can better account for the variation with respect to this additional metadata. On the other hand, contemporaneous work has demonstrated the use of contextualized embeddings, highly powerful word representations, in detecting semantic changes (Giulianelli et al. 2020). The next two chapters of this thesis expand on both these ideas and show how to incorporate them in answering the key question of this thesis *i.e.* of finding *who* is the leader of language change.

CHAPTER 6
SEMANTIC LEADERSHIP NETWORK INDUCTION FOR ABOLITIONIST NEWSPAPERS

6.1 Motivation

The previous chapter motivated to go beyond modeling *what* is semantically changing to *who* is leading or trailing semantic changes. One instantiation of this task, seen in chapter 5, is finding documents in a timestamped collection as semantic leaders: in this formulation, documents that use words with their changed meaning before others are deemed as leaders.

Many timestamped text collections, however, contain documents that are not produced by a single source; for example, in a collection of newspaper articles (as is the case in this chapter) the document sources are different newspapers.[1] Importantly, the document sources can be the underlying place of variation in the meaning of words at a given time; for example, two newspapers can use the same word for some issue with different ideological spins at the same time. The heterogeneity in the document sources can also extend to inferences about the lead-lag dynamics between the document sources; for example, a newspaper may introduce a new ideological framing of an issue that is picked up by other newspapers later.

Existing diachronic word embedding models are not fully equipped to account for the additional metadata of document sources in the modeling step. Moreover, none of the existing methods can quantify the lead of one source over another on a specific semantic change, or aggregate multiple changes together to get a summary view of the leader-follower structure between the document sources. This chapter fills this gap by proposing a computa-

[1] In the previous chapter also both the legal opinions and the scientific abstracts were produced by different courts and publishing venues respectively. However, this label on the document was not used in the modeling or the evaluation.

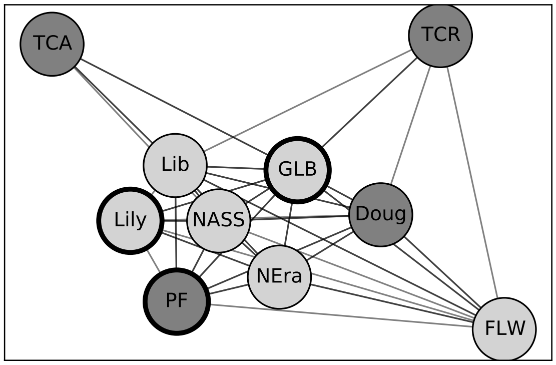
tional model to learn source dependent diachronic word embeddings from a timestamped text corpus. These embeddings are further used in a proposed metric of leadership between any pair of sources for a specific change. Finally, a statistical method is proposed to aggregate the lead-lag relationship between sources across many changes to induce a semantic leadership network, whose centrality analysis can reveal the overall leaders, followers or non-participants in the change.

The proposed methodology is applied to a collection of nineteenth century abolitionist newspapers. This is an important choice from a substantive standpoint since these newspapers played a crucial role in spreading information and shaping public opinion about the abolition of slavery and related social justice issues. However, until relatively recently, these newspapers — and the efforts of the white men who edited them — have dominated historical accounts of the abolitionist movement (Ernest 2004).[2] Though recent qualitative scholarship has highlighted the leadership role of the black press (e.g., Foster 2005; Gardner 2011; Fagan 2016; Spires 2019) and women (e.g., Sinha 2016; Jones 2009; Peterson 1995; Foreman 2009; Spires 2019) in the movement, quantitative evidence is still lacking.

By focusing on the *micro-structure* of the language in these archived newspapers, this work hypothesizes that aspects of social and political leadership — more specifically, the introduction of new concepts, the reframing of existing ones, and the advancement and circulation of both — can be identified in the appearance of individual words and the statistics that describe their lexical contexts. By connecting these words and contexts to the specific newspapers in which they appear, a new layer of evidence is offered about the emergence of certain social and political currents, the sources responsible for their transmission, and the networks that helped sustain them.[3] This novel computational approach offers a counterweight, however imperfect, to the forces of power that operate along the lines of both

[2] This is accentuated further because evidence from the past is gathered from what is archived which is lopsided in favor of white male editors (e.g., Trouillot 1995; Stoler 2009; Fuentes 2016)

[3] The code and some analysed data can be found at `https://github.com/sandeepsoni/semantic-leadership-network`. The supporting data for the analysis can be found at `https://doi.org/10.7910/DVN/EWYMFG`



Doug	Douglass Newspapers
FLW	Frank Leslie's Weekly
GLB	Godey's Lady's Book
Lib	The Liberator
Lily	The Lily
NASS	National Anti-Slavery Standard
NEra	The National Era
PF	The Provincial Freeman
TCA	The Colored American
TCR	The Christian Recorder

Figure 6.1: **Abolitionist newspaper network.** A spring layout of the ten newspapers in our collection, in which newspapers are positioned nearby in space to the extent that they influence each other (Fruchterman et al. 1991); the strength of the pairwise influence is also reflected in the darkness of the tie. Dark gray nodes represent members of the Black press; nodes with darker outlines represent publications edited by and/or aimed at women. For more details on the newspapers, see § 6.3.

race and gender, and that have overdetermined much of the historical scholarship on the abolitionist press to this point.

Citation Sandeep Soni, Lauren Klein, and Jacob Eisenstein. "Abolitionist Networks: Modeling language change in nineteenth-century activist newspapers". In: *Journal of Cultural Analytics* (2021)

6.2 Method: From timestamped text to aggregate influence networks

This section describes the method in more detail. A model to learn diachronic word embeddings from text is described in § 6.2.2. Further augmentation of the model to learn diachronic word embeddings that are conditional on the newspapers and which can be used to statistically measure the lead of one newspaper over other is described in § 6.2.3. Finally, a scheme is proposed to aggregate the lead-lag relationships of newspapers and quantify their influence in § 6.2.4 and § 6.2.5.

Mathematical notation. Define a document as a sequence of discrete tokens (words) from a finite vocabulary $\mathcal{V}$, so that document i is denoted $w_i = (w_{i,1}, w_{i,2}, \ldots, w_{i,n_i})$, with n_i indicating the length of document i. A corpus is a set of N documents, $\mathcal{W} = \{w_1, w_2, \ldots, w_N\}$. Each document is associated with a discrete time t_i and source label s_i. The time is created by binning the document timestamps into T bins; there are S sources, corresponding to each of the newspapers in the collection. The specific newspapers and time bins are described in § 6.3.

6.2.1 Word embeddings

The proposed approach builds on word embeddings, which represent each word in the vocabulary as a vector of real numbers (Deerwester et al. 1990; Turney et al. 2010). The embeddings are estimated by optimizing the classical **skipgram** objective, in which embeddings are parameters in a model of the probability of each word token conditioned on its neighboring tokens (Mikolov et al. 2013a). While the method is well known, some details are relevant to the innovative aspects of the proposed methodology, so they are reviewed briefly. Omitting the document index i, the skipgram objective is based on the probability,

$$P(w_{j'} \mid w_j) \propto \exp\left(\mathbf{v}_{w_{j'}} \cdot \mathbf{u}_{w_j}\right), \tag{6.1}$$

where $\mathbf{v}_{w_{j'}}$ is the "output" embedding of $w_{j'}$, and $\mathbf{u}_{w_j}$ is the "input" embedding of w_j. Each $\mathbf{v}$ and $\mathbf{u}$ is a vector, which is estimated by maximizing the product of these probabilities across all nearby pairs of tokens $(w_j, w_{j'})$. The result is that words that appear in similar contexts tend to have similar embedding vectors, which is the essence of the theory of distributional semantics (Harris 1954; Firth 1957; Lenci 2008). The input embeddings $\mathbf{u}$ are retained for further use, as in most of the prior applications of the skipgram embedding model to digital humanities research (e.g., Heuser 2017; Gavin 2018).

6.2.2 Identifying semantic changes

The skipgram embedding model is not equipped to use metadata such as times associated with the tokens. However, Bamman et al. (2014) proposed an extension to specialize the embeddings accounting for any discrete labels attached to each token. The core idea of their approach is to decompose the input embedding of a word as a sum of two embeddings — a base embedding and a residual embedding. Though originally developed to learn distinct embeddings of words based on geography, here it is applied to learn diachronic embeddings (see also recent work by Gillani et al. 2019). To do this, time is discretized into T bins, with $t_i \in 1 \ldots T$ indicating the time bin for document i. Then if w appears at position j in document i, the following input embedding is substituted in Equation 6.1.

$$\mathbf{u}_{w_{i,j}}^{(t_i)} = \mathbf{b}_{w_{i,j}} + \mathbf{r}_{w_{i,j}}^{(t_i)}, \tag{6.2}$$

where $\mathbf{b}_{w_{i,j}}$ is the base embedding of the word $w_{i,j}$ and $\mathbf{r}_{w_{i,j}}^{(t_i)}$ is the residual for time t_i. The base embedding represents the elements of meaning that do not change over time; the residual embeddings are specialized for the contexts encountered at each temporal epoch, making it possible to model semantic change. To ensure that the base embedding is used to account for the time-invariant components of meaning, the residual embeddings are regularized by adding an l_2 penalty to the skipgram objective, corresponding to the sum of squared values, $\sum_w^{\mathcal{V}} \sum_t^T (\mathbf{r}_w^{(t)})^2$. This encourages the estimator to choose values of $\mathbf{r}$ that are close to zero unless a strong temporal signal is observed.

To detect semantic changes, the focus is on finding terms whose embeddings change substantially over time. Among many possible ways to quantify change in embeddings (e.g. Hamilton et al. 2016a; Asgari et al. 2020; Gonen et al. 2020), the method from Hamilton et al. (2016) is applied, in which the words are ranked for semantic change by computing the difference in the similarity of their neighbors over time in the embedding space. The method is extended to account for multiple time bins by computing the differences

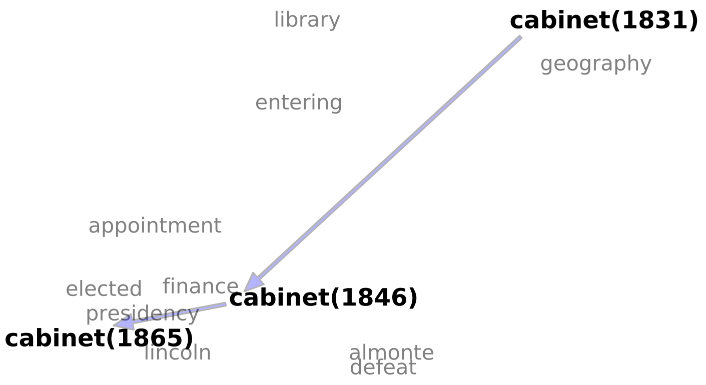


Figure 6.2: **Semantic change in *cabinet*.** The embeddings of *cabinet* and its near neighbors are projected to two dimensions using singular value decomposition. The embeddings of the target word indicate meaning shift from the architectural uses to political uses. This is reflected in the position of the embedding but even more so in the change in near neighbors.

between neighbors for all pairs of times in T for every word and picking the interval that has the maximum difference as the measure of semantic change in the word. Words are then ranked based on this measure, yielding an ordered list of changes as tuples of words and timestamps.

Figure 6.2 provides an example: the term *cabinet* enters the corpus being used in the (now) archaic sense to describe a small storage room, as in a cabinet of curiosities; but in the final years of the corpus, fully shifts to reflect its usage in the political sense, as in the context of a president's Cabinet (other changes pertinent to abolition are described below in the section § 6.4.1). As a measure of face validity of the proposed approach, many changes were found to refer to entities and concepts from the Civil War, such as *ferry* (later referring to *Harper's Ferry*), *confederates* (acquiring a specialized meaning after the formation of the Confederacy), and *battery* (from an assault to an array of guns). Changes referring to specific named entities were filtered at a later stage.

The proposed approach of modeling semantic changes has three main advantages. First, as a joint model of words and time, it does not require the computationally expensive post-alignment of the word embeddings that is common in standard approaches (e.g. Kulkarni et al. 2015; Hamilton et al. 2016b). Second, temporal embeddings can be learned even for words that emerge or disappear before the start or end of the time period respectively — these are just the base embeddings. Finally, the model is easily extended to incorporate other metadata about the text if available, as described in the next section.

6.2.3 Finding semantic leaders

Language changes are not adopted by everyone at the same time: for any change there are leaders, followers, and individuals who resist the change altogether. Moreover, being a leader for one semantic change does not imply semantic leadership for all changes (Raumolin-Brunberg 2006). It is of interest, therefore, to identify who led and followed in each semantic change. To do this, the diachronic embedding model is augmented to include an additional term for the *source* of each token — that is, the newspaper in which the token appears. The underlying input embedding representation can then be written,

$$\mathbf{u}_{w_{i,j}}^{(t_i,s_i)} = \mathbf{b}_{w_{i,j}} + \mathbf{r}_{w_{i,j}}^{(t_i)} + \mathbf{r}_{w_{i,j}}^{(t_i,s_i)}, \tag{6.3}$$

where $\mathbf{r}_{w_{i,j}}^{(t_i,s_i)}$ is the source-specific temporal deviation added to the temporal and atemporal components of the input embedding. As before, l_2 regularization is applied to the residual terms, which ensures that $\mathbf{r}_w^{(t,s)} = 0$ for words that are not used by source s at time t.

The fully conditioned embeddings are then used to assign a leadership score between a pair of sources for a given change (w, t_1, t_2), where $t_1 < t_2$ are the timestamps of a change in the meaning of word w. For each pair of sources, s_1 and s_2, two quantities are calculated: first, the similarity in meaning of the word for s_1 with s_2 at the two different times and, second, the similarity in meaning of the word for s_2 at the two different times. A measure of the lead of source s_1 over s_2 is the ratio of vector inner products,

$$\textsc{Lead}(s_1 \to s_2, w, t_1, t_2) = \frac{\mathbf{u}_w^{(t_1,s_1)} \cdot \mathbf{u}_w^{(t_2,s_2)}}{\mathbf{u}_w^{(t_1,s_2)} \cdot \mathbf{u}_w^{(t_2,s_2)}} \tag{6.4}$$

A higher score indicates more leadership, with a score of 1 corresponding to a baseline case of no leadership. For more intuition, consider a few cases:

- If s_1 is indeed leading s_2 on w, then s_1's use of the word with the new meaning should precede that of s_2. Equation 6.4 checks this precisely by comparing cross-source similarity (the numerator) with self-similarity (the denominator) in meaning across the two times. If the numerator is larger than the denominator, then s_1 is likely to have used the word with the new meaning before s_2 adopted it. Conversely, if the denominator is larger, then s_2 has not followed s_1 in its use of this word.

- If w does not change in meaning, then no newspaper should be deemed a leader.This is reflected in our measure because the numerator is unlikely to significantly exceed the denominator in this case.

- If a source s_i does not publish at time t_i, then $\mathbf{r}_w^{(t_i,s_i)} = 0$. This means that $\mathbf{u}_w^{(t_i,s_i)}$ will be identical to the global average at time t_i, and therefore s_i can be neither a strong leader or follower.

- Suppose w changes in meaning between time t_1 and t_2, but both s_1 and s_2 reach the new meaning in synchrony, as may happen if the change in meaning is caused by a sudden external event. In this case, s_1 and s_2 would both use an older sense of the word at t_1 and then switch to the newer sense at t_2. The numerator and denominator in Equation 6.4 will be approximately equal, resulting in a leadership measure close to 1.

The changes that took place over adjacent time periods are most likely to be indicative of inter-newspaper transmission. As a result, for each word in each successive temporal interval over the entire timespan, the lead of each newspaper over every other newspaper is calculated. Two newspapers (one leader, one follower) that produce the maximum lead score are identified. The pair is retained if it passes the threshold obtained from the randomization procedure described next.

6.2.4 Controlling for random noise and dataset artifacts

In any finite dataset, there will be spurious correlations, which may appear to be meaningful due only to random noise. In the dataset considered for this work, there is added risk due to two sources of heterogeneity of the data. As shown in Figure 6.3, some newspapers publish towards the beginning of the dataset (from 1831) and others towards the end (1861-1865); furthermore, the corpus overall contains more text published in the years leading up to the civil war. Second, the number and length of the articles published by each newspaper are widely divergent. While this heterogeneity reflects the reality of the historical trajectory of newspaper publishing at the time (e.g. Leonard 1995; Gross et al. 2010), it is potentially problematic: a statistical analysis of temporal trends might inherently focus on the newspapers that publish early and often as leaders, simply due to temporal precedence; similarly, newspapers that publish in greater volume may play an outsize role in determining the temporal word embeddings $\mathbf{u}_{w_j}^{(t_i)}$, so that the assessment of semantic leadership will be most sensitive to changes in the words that these newspapers emphasize.

While these issues can be partially mitigated by limiting the number of tokens per newspaper at each time step, additional controls are necessary due to the temporal heterogeneity of the data. Randomization is used to control for both structural heterogeneity and random noise. This is done by creating a set of $K = 100$ randomized datasets, in which word tokens are randomly swapped between newspapers. This preserves the structure of the dataset — each newspaper has the same number of tokens within each time period — but it breaks the link between individual newspapers and contextual word statistics. Any leadership relationships that are detected in such a randomized dataset must be attributed to either structural bias or random noise, because the word tokens have been assigned to individual newspapers at random.

To see how these randomized datasets help to control for noise and heterogeneity, recall that a leadership event is a tuple $(s_1 \rightarrow s_2, w, t_1, t_2)$, where s_1 and s_2 are newspaper sources, w is a word, and t_1 and t_2 are timestamps. For each such event, the leadership statistic from Equation 6.4 is computed in each randomized dataset, yielding a set of values $\{\textsc{Lead}^{(k)}(s_1 \rightarrow s_2, w, t_1, t_2)\}$ for $k \in 1 \ldots K$. These values are then compared with the score that was observed in the original, nonrandomized dataset. The final set of influence events is the set of tuples such that,

$$\textsc{Lead}(s_1 \rightarrow s_2, w, t_1, t_2) > \Phi_{.95}\left(\{\textsc{Lead}^{(k)}(s_1 \rightarrow s_2, w, t_1, t_2)\}_{k=1}^{K}\right), \tag{6.5}$$

where the function $\Phi_{.95}(S)$ selects the 95th percentile value of the set S; as a special case, $\Phi_{.5}(S)$ is the median of S. This procedure is based on principles from statistical significance testing (DeGroot et al. 2011), and allows to approximately bound the probability of each leadership score arising by chance.[4]

6.2.5 Aggregated semantic influence network

To generalize from individual leadership events, the events are aggregated into a semantic leadership network among the newspapers. Specifically, we an edge-weighted network, $G = (S, E)$, is constructed where, as before, S is the set of newspapers that form the nodes in the network, and every $e \in E$ is a weighted edge, denoted $e = (s_1 \rightarrow s_2, c_{12})$, with c_{12} indicating the number of words for which s_2 leads s_1.

The structure of this aggregate influence network is interrogated further using two graph-theoretic measures. The first is *pagerank*, which computes the overall centrality of each node in the network (Page et al. 1999); however, as shown later, this single measure fails to distinguish three cases of interest: newspapers that usually lead, newspapers that usually follow, and newspapers that generally do not engage with the rest of the network. For this reason, a more fine-grained analysis is performed using the *HITS* metrics of *hubs* and *authorities* (Kleinberg 1999). While details of these algorithms can be found in network analysis textbooks (e.g., Newman 2018), brief descriptions are provided here for

[4] Similar procedures which involve permutation have been used in network regression. For example, see Simpson (2001)

completeness.

Pagerank. The principle behind pagerank is that the importance of a node is a function of the importance of the nodes that link to it (its "backlinks"). In the context of this work, each edge $s_1 \to s_2$ is weighted by the number of innovations in which s_2 led s_1, and then this weight is normalized by the sum of weights on edges from s_1. This means that a newspaper has high pagerank if it leads other high pagerank newspapers, and particularly if it is the sole (or main) leader of those newspapers. Mathematically, the pagerank of node i is given by:

$$\textsc{Pagerank}_i = \alpha \sum_j \frac{A_{ij}}{\sum_k A_{kj}} \textsc{Pagerank}_j + \beta, \tag{6.6}$$

where A_{ij} is the weight on the edge from j to i (number of words for which i leads j), $\textsc{Pagerank}_j$ is the pagerank of node j, and α and β are additional free parameters.[5] This formulation is recursive: the pagerank of each node is computed in terms of the pageranks of the other nodes. The overall pagerank problem can be formulated as a system of linear equations (one for each node), and these equations can be solved simultaneously by factorization of the matrix A. Pagerank has been used in prior work in the digital humanities (Jockers 2012), and is closely related to eigenvector centrality, which is used more widely in the social sciences (Jackson 2010).

HITS. A limitation of pagerank is that it assigns only a single importance score to any newspaper, which fails to differentiate followers from newspapers that simply do not participate in ongoing semantic changes. A more fine-grained analysis technique, HITS, which stands for hyperlink-induced topic search, is another linear algebra-based network centrality algorithm, which decomposes central nodes into two complementary groups, **authorities** and **hubs**. Authorities are those nodes that are pointed to by high-scoring hubs, and conversely, hubs are nodes that point to high-scoring authorities. In the context of this

[5] We set $\alpha = 0.85$ and $\beta = 0.15/|S|$, as is typical of many applications of pagerank (see Berkhin 2005, for a discussion).

work, high scores as an authority and hub indicate the importance of a newspaper as a leader and follower respectively.

Mathematically, the authority and the hub score for a node i can be given by the paired equations,

$$\textsc{Authority}_i = \alpha \sum_j A_{ij}\ \textsc{Hub}_j, \tag{6.7}$$

$$\textsc{Hub}_i = \beta \sum_j A_{ji}\ \textsc{Authority}_j, \tag{6.8}$$

where A_{ij} is the weight on the edge from j to i (number of words for which i leads j), $\textsc{Authority}_i$ is the authority score of node i, $\textsc{Hub}_i$ is the hub score of node i.[6] Similar to pagerank, HITS also originated with the application of ranking webpages in a hyperlinked environment, but it has been increasingly used in information sciences (e.g. Jiang et al. 2012) and digital humanities (e.g. Sudhahar et al. 2015).

6.3 Data

This work analyzes a subset of nineteenth century newspapers digitized and hand-keyed by Accessible Archives.[7] Each newspaper in the archive consists of multiple issues, and each issue contains multiple articles. Each article is encoded as a unique `html` page. The text is extracted from each page using `beautifulsoup4` package in python, following a process similar to Klein et al. (2015) . The data is further processed using the methods described below.

[6] Unlike pagerank, α and β are not needed to be preset and are related to the leading eigen value of $A^T A$ and AA^T (Newman 2018).

[7] https://www.accessible-archives.com/

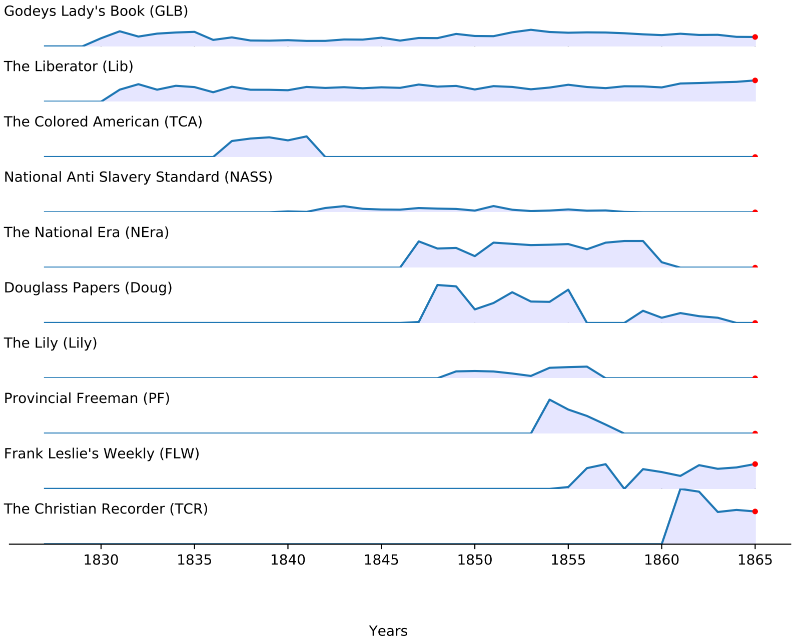


Figure 6.3: **Data distribution.** The distribution of data in Accessible archives, after further processing to remove duplication and fixing digital errors, that is used for analysis in the rest of the paper.

6.3.1 Contents and Metadata

For the purposes of this study, the corpus is limited to the period between March 1827, the date of the earliest newspaper in the collection, and December 1865, which marks the ratification of the Emancipation Proclamation. Most newspapers in this collection published weekly, though some titles are monthly publications. The temporal distribution of the newspapers in the collection is given in Figure 6.3, and a detailed description of the newspapers is found in Table 6.1. The collection is temporally skewed: the later years have more newspapers and more articles. This reflects a general trend in the rise in abolitionist newspapers published in the United States (Leonard 1995), as well as the influence of scholarship on the subject, which has focused on specific titles published in the final decades of this study, leading to their being digitized more quickly and more completely than other less-known works (Fagan 2016).

The focus of this study is on the abolitionist movement of the nineteenth-century United

Table 6.1: **Detailed description of all newspaper titles**. *Freedom's Journal* is included in identifying semantic changes but excluded from the subsequent analysis because its publication stopped very early in the timeframe that considered in this study.

Title	First Issue	Description
Freedoms Journal	1827	An abolitionist newspaper established by two Black clergymen, Samuel E. Cornish and John Brown Russworm. The earliest newspaper edited by and for Black Americans presently known.
The Liberator	1831	An abolitionist newspaper established by William Lloyd Garrison, the famed white abolitionist. It was known for its fiery rhetoric and is considered among the most influential papers of its day.
The Colored American	1837	An abolitionist newspaper established by Phillip A. Bell as *The Weekly Advocate*. Samuel Cornish, of *Freedom's Journal*, became its editor several months later, and initiated the name change. *The Colored American* was intended to serve as a national publication edited by and for Black Americans.
National Anti-Slavery Standard	1840	The official newspaper of the majority-white American Anti-Slavery Society. The author and abolitionist Lydia Maria Child was its first editor. It was intended to serve as a more moderate counterpart to *The Liberator*, and appeal to a white reading public.
The Douglass Papers	1847	A set of abolitionist newspapers edited Frederick Douglass, the author and orator who liberated himself from slavery. *The North Star* was launched in 1847, with Martin Delaney as co-editor. Douglass renamed it *Frederick Douglass's Paper* upon Delaney's departure in 1851. After a several year pause, Douglass launched *Douglass's Monthly* in 1859.
The National Era	1847	A more general newspaper, founded by Dr. Gamaliel Bailey, Jr. and aimed at a white audience. It supported the cause of abolition, and is most famous for being the first to publish *Uncle Tom's Cabin*.
The Christian Recorder	1854	A newspaper established by the African Methodist Episcopal (AME) Church. Once overlooked for its contributions to abolition, it is now understood as central to the movement for Black liberation in the nineteenth century.
Provincial Freeman	1854	An abolitionist newspaper established by Mary Ann Shadd, a Black abolitionist, educator, and later lawyer, who emigrated from the US to Canada. The *Freeman* is believed to be the first newspaper edited by a Black woman in all of North America.
The Lily	1849	The first known women's suffrage newspaper founded by Amelia Bloomer, who was white. Its initial emphasis was on the temperance movement, but it became more invested in women's rights over time.
Godeys Ladys Book	1830	A monthly women's magazine edited by a white woman, Sarah Josepha Hale, for the majority of its run. It was rarely overtly political. It is included in the corpus as a point of comparison with *The Lily*.
Frank Leslies Weekly	1855	A weekly magazine aimed at a white reading public. It offered a wide range of content, and is best known for its literary selections and coverage of current events. It is included in this corpus as a point of comparison to both the abolitionist and suffragist papers of the same era.

States. This is reflected in the titles included in the corpus, the majority of which explicitly identify as abolitionist newspapers. The corpus is notable for its relatively equal number of black-edited and white-edited abolitionist titles; this was among the reasons Accessible Archives was selected as the data source. Accessible Archives also provides information about the race and gender of the editors of each newspaper, as well as about the paper's intended audience; the identity of each editor and each newspaper's audience was confirmed with a second scholarly source, and this information was converted to use as additional metadata in the project. During this period, most newspapers were associated with a small group of editors (on the order of one to three people), which remained relatively constant over the newspaper's run, and who exerted a high degree of control over the newspaper's contents (Casey 2017).

The abolitionist titles were supplemented with additional newspapers linked to the women's suffrage movement, so as to enable a comparison between the two movements, which were often linked. A more general newspaper of that same era, as well as two monthly magazines are also included. The rationale for including these titles is to provide a baseline for any claims about abolition (or women's rights) that are made. While this dataset is far from exhaustive, it represents a meaningful "scholarly edition of a literary system," as Katherine Bode might describe it, one which captures some of the key titles in the scholarship on the subject along with known interlocutors and additional likely influences (Bode 2018). For more on the history of these newspapers, see Gross et al. (2010)

.

6.3.2 Data Processing

Digital error correction. While Accessible Archives' hand-keyed texts are far less error-prone than those digitized using optical character recognition(OCR) (Smith 2007), several types of errors were discovered in the course of our analysis. One common source of error in this collection is whitespace.[8] Such errors, also known as *word segmentation errors* or *spacing errors*, arise during both OCR and the post-digitization handling of the data (Kissos et al. 2016) — the latter being the source of the error in the Accessible Archives corpus — and result in the elimination of whitespace between words. This leads to out-of-vocabulary items like *senatoradmits* and *endowedwith*. The approach to correcting these errors is described in prior work (Soni et al. 2019a).

Deduplication. The collection also contains a number of articles that were reprinted verbatim from other newspapers (e.g. mission statements, notices, speeches) as well as many that were reprinted with minor modifications (e.g. advertisement campaigns, commodity price reports). While these reprinted articles are an artifact of the time (see Cordell et al. Forthcoming), their presence in the corpus is problematic for modeling, potentially resulting in false positive semantic changes. As a conservative heuristic, all articles that share a contiguous sequence of eight or more words with another article are removed. Though this heuristic can also remove a number of non-duplicate articles, it is still applied because it does not affect the distribution of the documents temporally or across newspapers, and keeps roughly 90,000 articles in the corpus.

Data and experimental setup. For all the analysis, the corpus is divided into ten equal time intervals.[9] Newspapers that do not contain at least 500 articles throughout their tenure in the period of interest are ignored. To maximize the number of newspapers that meet this condition, several smaller newspapers that share an editor are grouped together. Namely, the THE NORTH STAR, FREDERICK DOUGLASS'S PAPER, and the DOUGLASS MONTHLY are consolidated into a single source labeled as DOUGLASS PAPERS, because

[8] Other digital errors (e.g. first or last letter of a word getting stripped) also occur in the corpus but are not systematic.

[9] With 5 and 20 time intervals, the discovered semantic changes were very similar, with approximately 90% of all changes preserved across all three settings. However, increasing the number of intervals made the residuals more difficult to compute, while decreasing the number of intervals made it difficult to isolate the precise time of changes. Consequently, the number of intervals is set to ten as a tradeoff.

each was edited by the black abolitionist and orator Frederick Douglass; THE WEEKLY ADVOCATE is grouped with THE COLORED AMERICAN, since the former changed its name to the latter several months into its publication run.

Vocabulary and hyperparameters. All the text is lowercased before processing. The vocabulary is limited to the $50,000$ most frequent words in the corpus. The embedding size for all the models is set to 100 dimensions. The window size to determine the context words is set to 5 in both directions. The ℓ_2 regularization penalty is set to $1e^{-4}$.

6.4 Results

The proposed method offers insights about the evolution of conceptual terms related to abolition, as well as about the specific newspapers that contributed to the spread of the terms' changed meanings. By considering these terms in relation to others identified by the model as associated with particular newspapers, an argument can be made about the nature of semantic leadership and its relation to social and political influence within the newspaper corpus. This section also discusses the significance of the overall influence network generated by our model, which confirms aspects of our previous understanding of the movement's intellectual leaders and followers, while illuminating additional pathways of influence among some of the understudied newspapers in our corpus. Taken together, these results help to establish the validity of the proposed method. The findings that result from each phase of the methodological pipeline are presented next, building towards the conclusion about women's leadership and the role of the black press.

6.4.1 Semantic changes

The first step in the proposed method identifies terms that undergo semantic change. Illustrative examples are documented in Table 6.2. The word *equality*, in its earliest usage in the corpus (1827-1831), seems to express the idea of equality in a Lockean sense. It is

Table 6.2: **Examples of semantic changes**. The terms are detected to have changed most in meaning from the earlier timespan to the later.

Word	From	Until	Earlier examples	Later examples
cabinet	1831-1835	1861-1865	*a **cabinet** of antiquities* *the secret **cabinet** on the right*	*president and his **cabinet*** *members of the **cabinet***
equality	1827-1831	1853-1857	*liberty and **equality** to all* *blessed with liberty and **equality***	*the **equality** of the races* *an **equality** with the whites*
freedom	1827-1831	1857-1861	*the spirit of **freedom** is marching* *establish their **freedom** from being sold*	*opposition of slavery to **freedom** continues* *colored man for **freedom** and self-government*
rights	1841-1845	1861-1865	*inalienable **rights** of man*	*advocate the **rights** of the black man* *the **rights** of women*
justice	1831-1835	1861-1865	*legislature or the court of **justice*** *james madison and chief **justice** marshall*	*one simple act of **justice** to the slave* *emancipation as an act of **justice** and humanity*
immediate	1827-1831	1849-1853	*the **immediate** legislation of congress* *the **immediate** extinction of slavery*	*hostile to **immediate** emancipation* *safety of **immediate** unconditional emancipation*
fight	1827-1831	1849-1853	*you are paid to **fight*** *a **fight** took place*	*profess to **fight** for liberty* *determination to **fight** till the last*
service	1831-1835	1861-1865	*first military **service** was in corsica* *their **service** in the army*	*gallant **service** at fort wagner* ***service** was held in the church*
aid	1831-1835	1861-1865	*offers of **aid** from colonizationists* *gave them **aid** and tyrants fled*	*state fugitive **aid** society* *liberal offers to **aid** the cause of freedom*
growing	1835-1839	1861-1865	***growing** up in shame and poverty* *extreme want **growing** out of the toil*	*republic is **growing** darker everyday* *stole the **growing** light of dawn*
writing	1827-1831	1853-1857	*coloured adults in reading, **writing**, arithmetic* *reading or **writing** the letters seem*	*the **writing** is blotted in many places* *full of **writing** in a round-text hand*
hoped	1841-1845	1845-1849	*it is **hoped** that every newspaper* *it was **hoped** however that the friends of temperance*	*promised and **hoped** for but homeless future* *he now **hoped** to have peace*
courage	1827-1831	1849-1853	*possessed of both sense and **courage*** *wit and **courage** amongst all*	*renewed **courage** instead of the hopeless feeling of banishment* *personal **courage** and devotion*

nearest in meaning to other key terms in the Enlightenment discourse of natural rights, such as *liberty*, *rights*, *wisdom*, and *mankind* — the same terms that were commonplace at the time of the nation's founding, and into the early years of the republic. Over time, however, the usage of the term becomes more connected to questions of the practice of democracy. Although it retains some of its original associations, by 1857 *equality* also becomes closely associated with the idea of *self-government* issues of *suffrage*, and the specific *guarantees* of the ideals expressed in nation's founding documents. While this shift from abstract to concrete is often attributed to the authors of the nation's founding documents, as they attempted to translate political philosophy into a governing structure, it is interesting to observe that in the abolitionist press (and related publications), the shift happens several decades later. While the changing meaning of *equality* does not offer a definitive answer to the underlying reasons behind the change, it helps point to how the changing meanings of individual words, as seen in the corpus, can index larger ideological changes and debates.

The term *freedom* follows a similar, if more chronologically consistent shift from abstract to concrete. It enters the corpus with general associations to the nation's foundational ideals. Its near neighbors include words like *humanity*, *people*, and *country*. By the time of beginning of the Civil War, however, the associations of the term have become more specific, including references to both *liberty* and *slavery*, as well as terms that reference arguments about the *rights* and *guarantees* of the nation that should ensure freedom, as well as the *institutions* that should enforce it. The term *rights* also moves from a more abstract to a more concrete meaning, tracking both *freedom* and *equality* — terms which appear in its list of near neighbors, along with terms like *humanity*, which reinforce a sense of the term that is broadly applied. Over the second half of the corpus, however, the term narrows in its application to questions of *citizenship* and *suffrage*. As with the term *equality*, knowledge of which newspaper was most responsible for this narrowed sense of the term might tell something about who was responsible for shifting the conversation about *rights* and their application.

A final term, *justice*, is worth exploring in this context. Like *freedom* and *rights*, this term also follows a trajectory that is consistent with historical scholarship on the subject, albeit one that moves in the reverse direction from concrete to abstract. It enters the corpus with narrow associations to the legal system, as evidenced by near neighbors including *judges*, *trial*, *offense*, and *crime*. But by the final years of the corpus, the term has significantly expanded, commanding a more ideological frame. Its associations at the onset of the Civil War include the terms *liberty*, *equality*, *rights*, and *oppression*, which tracks arguments about how justice transformed into a much more powerful concept over the course of the nineteenth century as a result of the criminal justice movement. With respect to the current corpus, this transformation is reinforced by the appearance of the terms *universal*, *humanity*, *citizenship*, and *nation*, which suggest the success of those who advocated for criminal justice in expanding not only legal protections, but also ideas about what *justice* properly entailed.

Taken together, these terms demonstrate how changes in near neighbors, coupled with information about which newspapers were most responsible for those changes, can point to a new understanding of (or, alternately, enhance existing knowledge about) the abolitionist movement and its conceptual undercurrents.

6.4.2 Semantic leadership

Of the changes discussed in the § 6.4.1, two — *justice* and *rights* — are attributable to specific newspapers. In terms of *justice*, interestingly, the newspaper most closely associated with the more ideological conception of the word was not an abolitionist newspaper at all, but rather, the women's suffrage newspaper, THE LILY. More interesting still, it is THE LIBERATOR — the long-running abolitionist newspaper with a reputation for its impassioned arguments against slavery — which is identified as the most significant follower in this new usage. Could it be that THE LIBERATOR, edited by a white abolitionist, William Lloyd Garrison, was not as ideologically innovative as the historical narrative would lead us to believe? This line of inquiry is continued below.

As for *rights*, the method identifies the two more general newspapers, both intended for a white readership, as the leader-follower pair. In the period between 1853 and 1861, THE NATIONAL ERA is the most responsible for narrowing the scope of the conversation about rights from abstract and ideological to legal and concrete; and it is FRANK LESLIE'S WEEKLY, a weekly magazine which also published a wide range of content, which most closely followed this trend. That the conversation about rights is bounded by white, general-audience newspapers suggests that this concept, with its anchor in political philosophy, may hold less relevance to the abolitionist press, or to women's suffrage newspapers, than concepts involving ideas about morality, humanity, or other more expansive understandings about the implications of the persistence of slavery.

Beyond the analysis of individual words and the newspapers that lead in their changing meanings, it is also possible to examine any two newspapers as leader-follower pairs. Comparing two newspapers that published at the same time as each other, and were known to be in conversation — for example, THE COLORED AMERICAN, a newspaper edited by a black editorial staff which circulated among a predominantly black audience, and THE LIBERATOR, mentioned above, edited by William Lloyd Garrison, who was white, and which circulated among a mixed, if predominantly white audience — it is found that THE COLORED AMERICAN leads on the terms *immediate* and *fight*. While not explicitly political terms, they do suggest a tone of urgency that might surround an argument for liberation, one which helps to further dismantle the narrative of THE LIBERATOR as the most radical of the abolitionist papers, and in the process, provides additional evidence to support a claim that THE COLORED AMERICAN, led by black abolitionists, deserves more of the credit for accelerating the fight against slavery.

When considering how newspaper pairs and the words that connect them can open up new research questions, we might consider the leader-follower relationship between THE CHRISTIAN RECORDER, the official newspaper of the African Methodist Episcopal

(AME) Church, and Godey's Lady's Book, a magazine aimed at white women. The Christian Recorder is among the several publications of the AME Church, including The Christian Recorder that have been posited as key sources in the emergence of a Black print culture (Foster 2005; Fagan 2016), and yet The Christian Recorder stands far to the side in the induced influence network, neither leading nor following many other newspapers. This suggests that the conversations taking place in that paper were disconnected from the others — in terms of language and therefore abolitionist thought. A prior thematic analysis confirms that the articles did indeed consist mostly of religious content (Klein 2020). Yet the proposed model detects that the paper *does* lead the rest of the network on the terms *service* and *aid*. These terms clearly relate to core tenets of Christianity, but is interesting to note that the main follower is not another abolitionist paper but instead a white women's magazine. This suggests an unexpected new line of inquiry: how Christian benevolence, often framed as the contribution of white women's involvement in the abolitionist movement, was instead rooted in the black church.

Thus far, only individual newspaper pairs and the specific words which connect them are considered. But these words can also be analyzed in aggregate. The proposed method in this study identifies 435 semantic leadership events, containing a mixture of political terms and activist language, as well as other words that are not easily connected to any particular political or ideological stance. In fact, words like *growing* and *writing*, *hoped* and *courage* — words which suggest generalized movement, motion, action, and emotion — emerged with some of the highest leadership scores. This finding suggests that the influence detected by our model might be better understood at the level of what might be described as *discourse*, rather than at the level of individual words.

Looking at the aggregated counts of leader-follower pairs confirms this hypothesis: the relationship between The Liberator and The National Anti-Slavery Standard rises to the top of the list, with 32 words (of the 435) on which The Liberator leads and The National Anti-Slavery Standard follows. This is a valuable if unsurpris-

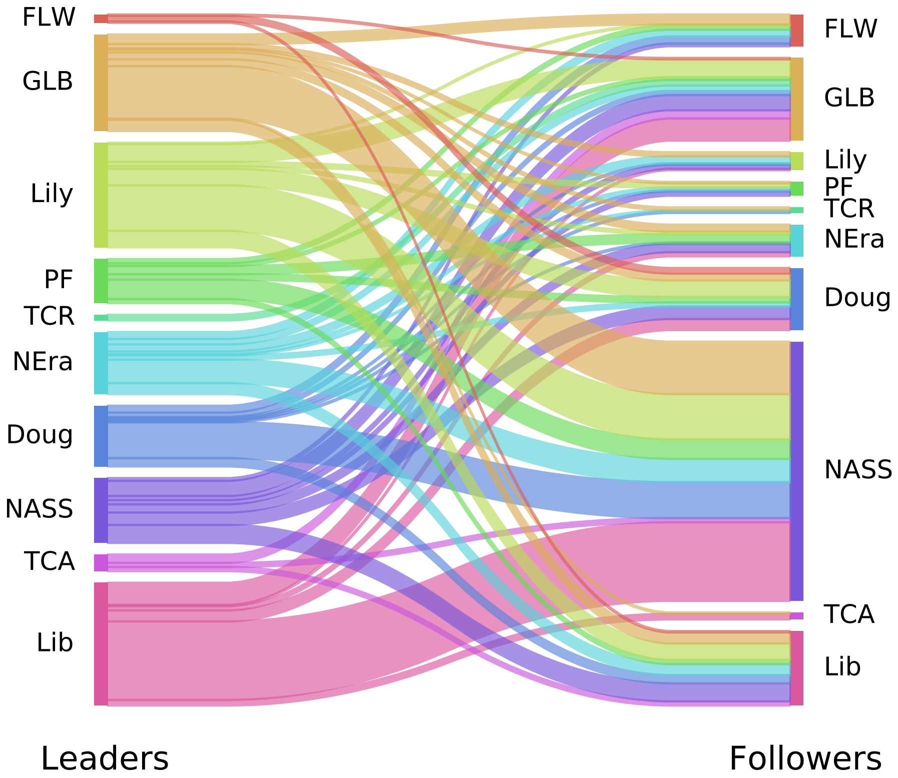
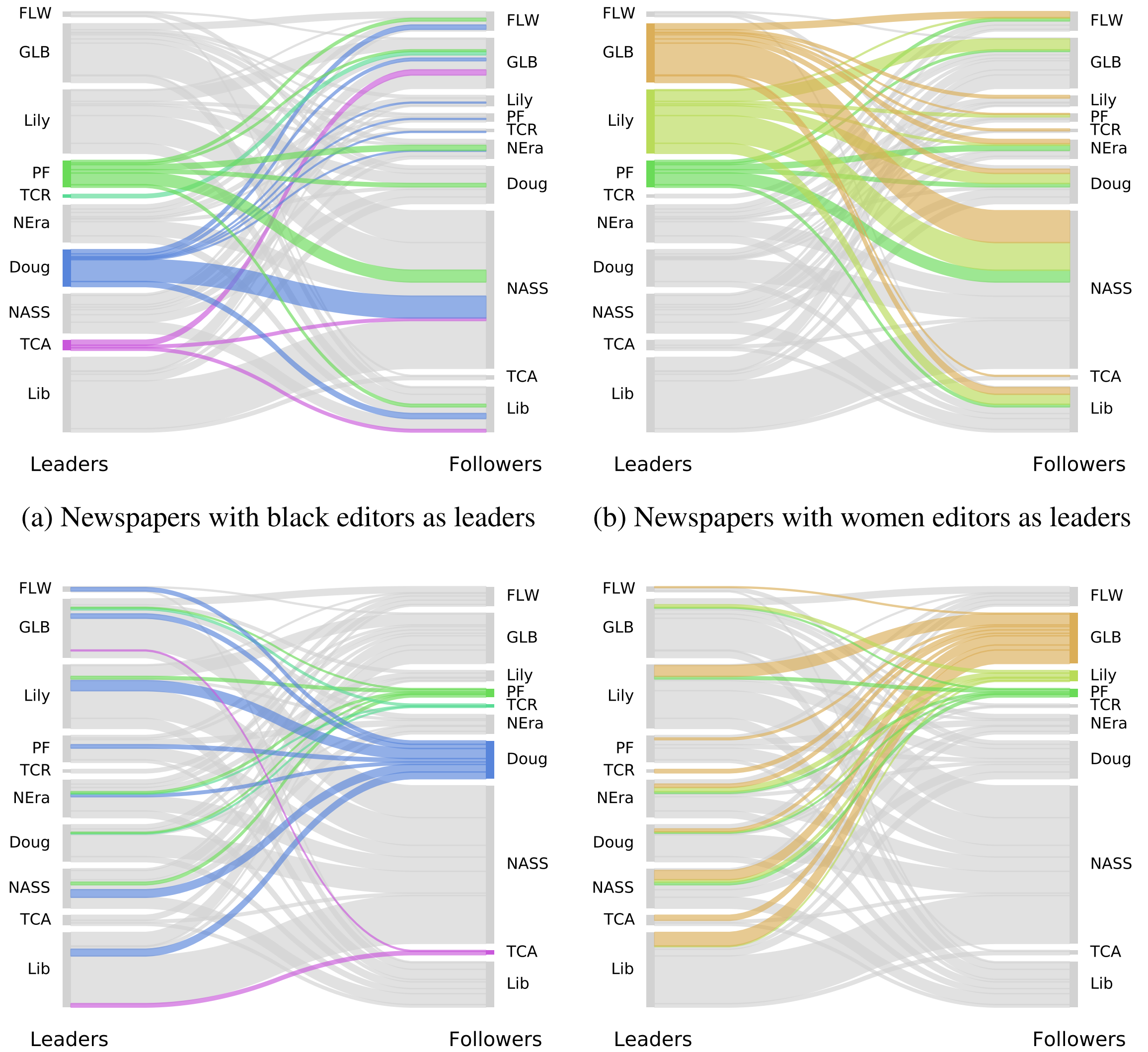


Figure 6.4: **The leader-follower newspaper pairs**. The leading newspapers are shown on the left and the trailing newspapers on the right. Each newspaper is shown as a rectangular patch with height proportional to the number of words for which it is considered a leader (or follower). The thickness of each stripe connecting two newspapers is proportional to the number of words between the newspapers that the stripe connects.



(a) Newspapers with black editors as leaders

(b) Newspapers with women editors as leaders

(c) Newspapers with black editors as followers

(d) Newspapers with women editors as followers

Figure 6.5: **Focused lead-lag relationships** Highlighting the dyads of newspapers that had black editors (left) and women editors (right). The top row shows the dyads in which these newspapers lead, while the bottom row shows the dyads in which they trail. For more details on the editorial coding of newspapers, see § 6.3.1.

ing result, since of all of the newspapers in the corpus, historical circumstance suggests these two titles *should* be in close dialogue with each other, as they were both published by the American Anti-Slavery Society. The relationship with the second highest count, between GODEY'S LADY'S BOOK and the STANDARD, is more surprising: the STANDARD was launched in no small part to bring women around to the abolitionist cause, and a woman — the white abolitionist and author Lydia Maria Child — edited the paper in its early years. The topical analysis performed by Klein (2020) revealed that the STANDARD did indeed contain more thematic content related to women, but the identification of the STANDARD as a follower of GODEY'S LADY'S BOOK provides new evidence about the centrality of women's issues to STANDARD. Intriguingly, the STANDARD also follows THE PROVINCIAL FREEMAN, an abolitionist newspaper edited by a black woman, in outsized degree to the FREEMAN's far smaller circulation. Edited by Mary Ann Shadd (later Carey), the FREEMAN was known, like THE LIBERATOR, for its uncompromising editorial tone (Rhodes 1998; Casey 2019). As a result, it struggled to find a wide readership (Rhodes 1998). It was also published out of western Canada, which further limited its reach. And yet there is evidence of its influence reaching quite far indeed, through the pages of the widely-circulated NATIONAL ANTI-SLAVERY STANDARD, which adopted aspects of its discourse. Notably, the FREEMAN also holds sway over the newspapers edited by Frederick Douglass, which have dominated accounts of the black press. Taken together, these aggregated leader-follower pairs point to how a quantitative analysis conducted at the level of changes in individual words, and subsequently subjected to an analysis in terms of leadership, can help shift the narrative about the abolition and those responsible for shaping its discourse.

6.4.3 Semantic Leadership Network

It is also possible to aggregate the leader-follower pairs and their changes all together in order to form a directed, weighted network. Figure 6.4 displays the leader-follower relationships among all of the newspapers considered in this study. In this diagram, all of the newspapers in the corpus are listed in order on both sides. A chord that emerges from a newspaper's name on the left indicates that it leads on a single word, while one that connects on the right indicates that the newspaper follows on that term. Thus, the vertical length next to any particular title indicates the number of words that the newspaper leads on or follows, respectively. Examining the titles associated with the highest number of leading and following words confirms two known findings: first, that THE LIBERATOR holds broad influence over many of the other newspapers, as intimated above; and second, that the NATIONAL ANTI-SLAVERY STANDARD tends to follow much more than it leads. This reflects the rationale for launching the paper, which was to provide a more moderate publication than THE LIBERATOR so as to appeal to a broader coalition of potential supporters. Recall that both newspapers were published by the AMERICAN ANTI-SLAVERY SOCIETY.

Just below those top-level leaders and followers, the most prominent leaders are THE LILY and GODEY'S LADY'S BOOK. Both newspapers (mostly) edited by and written for white women; neither is an abolitionist newspaper, so it is surprising that they are found to lead in a corpus are centered on abolition. While more research is required in order to understand the source of this influence, we can learn a bit more by examining the relationships among them, as pictured in Figure 6.5. Two things are seen: first, that the influence of these publications can be detected in every single newspaper in the corpus; and second, that THE NATIONAL ANTI-SLAVERY STANDARD and Frederick Douglass's various papers claim the majority of their influences from these women's publications.

When considering the sub-network of newspapers associated with the black press (Figure 6.5), the influence of these papers is again observed to be distributed across the corpus. Two newspapers dominate this sub-network: Frederick Douglass's newspapers and THE PROVINCIAL FREEMAN. But it is worth noting the titles that are far less embedded in the network: THE COLORED AMERICAN and THE CHRISTIAN RECORDER, which rarely lead *or* follow. This does not imply that they were not influential in their own right, but it does imply that their discourse was not connected to the discourse of this larger, white-dominated print network. It is quite possible — as has been suggested in the qualitative scholarship on the subject — that the conversations documented in these papers stood apart from the dominant, white-led abolitionist discourse. Future research might examine these newspapers for additional evidence of what the conversation about abolition looked like in a predominately black print sphere.

6.4.4 Network centrality

Finally, to better understand the holistic roles of each newspaper in the network, more sophisticated network analytic methods of *pagerank* and *HITS*, are performed. Pagerank is based on the number of terms on which each newspaper leads, but which especially rewards newspaper who lead other leaders. High pagerank newspapers can therefore be viewed as centralized consolidators of influence. Many of the titles previously discussed exhibit high pagerank scores (shown by node size in Figure 6.6), including The National Anti-Slavery Standard and The Liberator. Again, Godey's Lady's Book and The Lily, as well as The National Era, are seen suggesting that these papers each played a central role in consolidating the abolitionist discourse.

The HITS metrics (Hub and Authority scores, shown by position in Figure 6.6) help to clarify this relationship by differentiate several distinct groups, as well as some outliers. Once again, National Anti-Slavery Standard stands out with the largest Hub score by far, indicating its position as a fast follower (but not originator) of semantic change. Conversely, The Liberator stands out with the largest Authority score, indicating its position as a leader. A cluster of newspapers — Godey's Lady's Book, the Douglass Papers, as well as the The Liberator — scored high on both metrics, indicating their active participation as both leaders and followers of semantic change.

Returning to the initial focus on leadership, there are two titles with high Authority

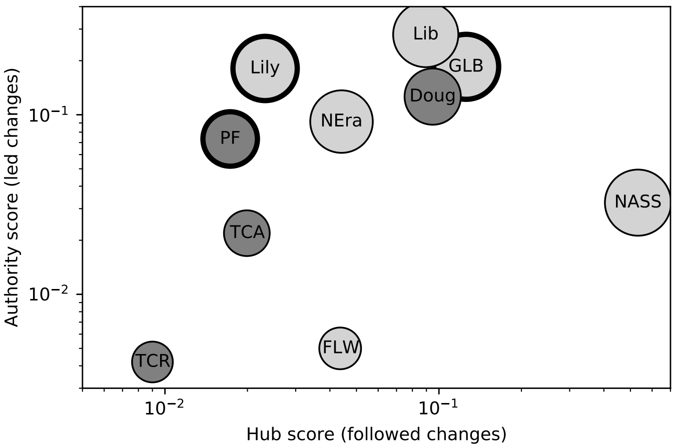


Figure 6.6: **Aggregate network statistics.** For each newspaper, the hub score roughly corresponds to how often it followed ongoing semantic changes, and the authority score to how often it led; node size corresponds to Pagerank. Node abbreviations are given in Figure 6.1. As in Figure 6.1, newspapers with black editors are shown in dark gray, and newspapers that were predominantly edited by and/or aimed at women have darker outlines.

scores but low Hub scores, indicating that they lead but rarely follow. These are the two titles, THE LILY and THE PROVINCIAL FREEMAN, both of which were edited by women. That THE LILY, women's suffrage paper, is so influential as a leader in this network of abolitionist newspapers is a finding that points to the need for additional research, as it might suggest a potentially revisionist narrative to the story that has the white women's suffrage movement departing from — and, at times, explicitly opposing — the abolitionist cause.

That the PROVINCIAL FREEMAN is the second title on this list is more intriguing still. Upon announcing her departure from the paper, Mary Ann Shadd, the newspaper's founder and editor, lamented that "Few, if any females had had to contend against the same business" that she had faced, referring to the criticism she had received as a result of her hard-hitting editorial style (Shadd 1855). She despairs at the lack of recognition that her paper had commanded, in spite of her having "broken the Editorial ice" for "colored women everywhere" by serving as the first black woman to edit a newspaper in all of North America. And yet here in this chart is evidence of the influence that Shadd indeed commanded through her editorial work.

6.5 Validity

An alternative proposal for the analysis in this chapter is to use word embeddings to distinguish between word senses and then retain only the "new" senses of the words to study their adoption by the newspapers — an approach used in contemporary research (Li et al. 2021). Such an argument can also be made for the study described in chapter 5. However, instead of discarding the word embeddings after performing word sense induction, we retain the word embeddings because they capture more high-dimensional information about the usage of the word; a further counterargument is that it is better to retain the word embeddings in the downstream task rather than make design choices with them in the upstream pipeline. Moreover, distinguishing the word senses does not help in identifying progressive usages, especially with non-contextual embeddings, as was needed in chapter 5.

Another alternative proposal for the network analysis is to use other measures of centrality, such as degree, closeness, or betweenness centrality, in place of pagerank and HITS. The closeness and betweenness centrality are not suitable for the problem at hand since they are path-based measures. In contrast, pagerank and HITS are spectral measures on the connectivity of the network. Pagerank and HITS are also versatile, simple to calculate, easy to interpret, and have found applications in text retrieval (Mihalcea et al. 2011). An argument can be made to the use of degree centrality which is also simple and interpretable. It should be noted that degree centrality is used in our visual analysis of the newspapers in Figure 6.4 and Figure 6.5. Moreover, pagerank and HITS emphasize finding newspapers as leaders or followers on the basis of finding other newspapers as leaders and followers, which degree centrality does not account.

6.6 Summary

The archive of abolition will always be bound by the historical forces that contributed to its creation — what Michel-Rolph (Trouillot 1995) has described as the "silences" of the past.

But these silences have not stopped scholars from attempting to derive meaning from the documents the archive does contain. On the contrary, the constraints of the archive have prompted a range of scholars to develop powerful new methods for the writing of history, methods which can expand our knowledge of the past.

This study contributes to these efforts. By considering the meanings of words over time, their shifts in meaning surface the conceptual undercurrents of the abolitionist movement of the nineteenth-century United States. Similarly, the role of individual newspapers with respect to these changes is characterized by the extent to which they led or followed other newspapers in the adoption of these changes. In some cases, these semantic changes are clearly at the core of the concerns of the abolitionist movement, as in the case of terms like *justice* and *rights*. Indeed, such examples are so central to the narrative of abolition that they merit close individual analysis through non-computational techniques. Yet the bulk of the semantic changes uncovered by our method are not of this type — they are "everyday" words like *busy* and *bold*, *ears* and *arms*. The story of these changes is best understood through aggregation, by identifying newspapers that consistently led their neighbors in the adoption (and in some cases, instigation) of these shifts. This aggregate view confirms some existing intuitions: for example, THE LIBERATOR, known for its role on the movement's vanguard, is found at the center (Figure 6.1), while FRANK LESLIE'S WEEKLY, which was intentionally less political, is placed at the periphery. Aggregation also enables the use of network statistics to distinguish classes of participants: leaders, like THE LILY; fast followers, like NATIONAL ANTI-SLAVERY STANDARD; and outsiders, like THE CHRISTIAN RECORDER. The aggregate view also proposes some more unexpected results, such as the prominence of the GODEY'S LADY'S BOOK, a magazine that sought to position itself apart from the political fray.

From a methodological perspective, a key intervention in this work is the use of randomization to ensure that are findings are robust. This was necessary for two reasons. First, although our dataset is large, we are interested in relatively rare phenomena — the appearances of individual words in specific linguistic contexts by each newspaper at several different points in time. In any finite dataset, rare phenomena may give rise to spurious correlations which are not represent of meaningful underlying trends. Second, our dataset is heterogeneous: because some newspapers publish earlier and more often than others, temporal methods for detecting influence will inherently favor these newspapers. Randomization addresses both of these issues by comparing the quantitative results against the distribution of results obtained from a large set of alternative datasets in which no semantic leadership is possible by construction. This methodology of control by randomization is broadly applicable, and is especially applicable to analysis of temporal phenomena, which is difficult to validate using traditional hypothesis testing (e.g., Dubossarsky et al. 2017; Dubossarsky et al. 2019).

CHAPTER 7
MODELING SEMANTIC LEADERSHIP USING CONTEXTUAL WORD EMBEDDINGS

7.1 Motivation

Diachronic word embeddings can be used to detect semantic changes in text, as demonstrated in the previous two chapters. Yet, diachronic embeddings word embeddings that are build from static embeddings (e.g., Mikolov et al. 2013a; Pennington et al. 2014) are limited in representation capacity since they map words into single vectors. To illustrate this point, consider the following hypothetical set of timestamped sentences containing the term *planning* in the corpus of papers from the ACL anthology. Assume for the sake of simplicity that t_1 and t_2 are the two timestamps.

t_1: These findings are informal and we are *planning* further exploration with a larger corpus.

t_1: While global *planning* is largely language independent, local *planning* can be language dependent.

t_2: The research uses *planning* for sentence-level natural language generation.

The term *planning* is used in the conventional sense in the first sentence and in a much narrower technical sense in the remaining two sentences. There is even a subtle distinction between the two senses of the term in the second and third sentences: the second sentence references the term more abstractly whereas the third sentence is specific to a particular NLP task. Models described in the previous chapters assign the same type-level embedding to the two tokens of *planning* at time t_1. This single vector representation interpolates between the two senses of the word but collapses all the contextual information that is key to distinguish the senses. This problem extends to detecting semantic changes when multiple senses of the word evolve simultaneously such as for the term *planning* where the conventional sense of the word still exists and the technical sense of the word is undergoing an evolution from abstract to specific (also see more example scenarios in Shoemark et al. 2019). Another issue with aggregating multiple senses into one representation is it fails to capture the typically intermediate polysemous stages in word meaning evolution (Hooper 1991).

Recent advances in language research have given rise to large self-supervised language models (e.g., Peters et al. 2018; Devlin et al. 2019; Yang et al. 2019; Raffel et al. 2020) that produce contextual word embeddings — token-level representations of words that depend on their sentential context (Smith 2020). Contextual word representations can distinguish the different senses of the word; consequently, contextual word embeddings have been used in detecting semantic changes in text (e.g., Giulianelli et al. 2020; Kutuzov et al. 2020). A further advantage that contextual word embeddings have over static embeddings for semantic change detection is that words entering or exiting the vocabulary at some intermediate time can be trivially represented; models such as the one described in chapter 5 can not handle this case, whereas the one described in chapter 6 do so at the expense of additional parameters in the model that scale linearly with both the size of the vocabulary and the granularity of the timescale.

However, similar to semantic change models that detect words that change, the contextual semantic change models are not fully equipped to detect the leaders of semantic change. It is also unclear whether the linguistic leadership calculated using the more powerful contextual word embeddings is related to other established notions of influence such as citations in the case of the research domain. The research question for this chapter is:

RQ Can contextual word representations be used to investigate the connection between semantic leadership and other explicit markers of influence such as citations?

This chapter proposes the use of contextual embeddings for semantic change detection:

a metric based on mahalanobis distance is proposed to account for variation in the contextual embeddings that encourages the detection of smooth monotonic changes, as opposed to bursty changes. A transition point for any word, that is the earliest approximate time of meaning shift is obtained before a subsequent step to mark every instance of a semantic change as "new" (post-transition meaning) or "old" (pre-transition meaning) usage. Finally, HP models (from chapter 3) are used to induce a semantic leadership network, which places the framework of inducing a network from semantic changes (as discussed in chapter 6) on a more probabilistic footing. In an application of this overall methodology to semantic change propagation in ACL venues (conferences, workshops, journals, *etc*), the semantic leadership network shows inverse correlation with the citation network between the venues, though more rigorous analysis still remains to be done.

7.2 Method

This method proposed in this chapter is comprised of four steps: (1) contextualized embeddings to detect semantic changes; (2) labeling usages of a semantic change by a classifier as "old" or "new"; (3) subselection of new usages as input to an HP; (4) estimation of parameters from the HP which are then used to induce a leadership network. A schematic of the entire pipeline is shown in Figure 7.1. A detailed description of each of these steps is as follows.

7.2.1 Setup

The setup is similar to the previous chapter. Let a document be a sequence of discrete tokens from a finite vocabulary $\mathcal{V}$, so that document i is denoted $W_i = [w_i^{(1)}, w_i^{(2)}, \ldots, w_i^{(n_i)}]$, with n_i indicating the length of document i. A corpus is similarly defined as a set of N documents, $\mathcal{W} = \{W_1, W_2, ..., W_N\}$. Let $E_i = [\boldsymbol{e}_i^{(1)}, \boldsymbol{e}_i^{(2)}, \ldots, \boldsymbol{e}_i^{(n_i)}]$ be the sequence of non-contextual embeddings of every word in document W_i. Each document is also associated with a discrete time $t_i \in \mathcal{T}$ and a source label $s_i \in \mathcal{S}$. Further, let $c(\cdot)$ be a contextualizing model, which takes non-contextual representations of a word $w_i^{(k)}$, $\boldsymbol{e}_i^{(k)}$, and its surrounding context, the left context $E_i^{(<k)}$ and the right context $E_i^{(>k)}$, to produce a contextual word representation $\boldsymbol{h}_i^{(k)}$ for the word. Let $\textsc{mean}(\cdot)$ and $\textsc{var}(\cdot)$ be elementwise mean and variance functions for the embeddings. Let $\textsc{diag}(\cdot)$ be a function that takes a vector as input and converts it into a diagonal matrix. Let d be the size of the embeddings.

7.2.2 Semantic change detection

The first step in the pipeline is to detect semantic changes. This is done by first obtaining the contextual embeddings for every token in the corpus $\mathcal{W}$ using the contextualizing model as follows

$$\boldsymbol{h}_i^{(k)} = c(\boldsymbol{e}_i^{(k)}, E_i^{(<k)}, E_i^{(>k)}), \forall i, k \tag{7.1}$$

Next, all the contextual embeddings for every candidate semantic change $v \in \mathcal{V}$ are collected and sorted by time, *i.e.*, let $\boldsymbol{V} = [\boldsymbol{h}_i^{(k)} | w_i^{(k)} = v, \forall i, k; t_i \leq t_{i+1}]$. Thus, $\boldsymbol{V} \in \mathbb{R}^{m \times d}$ is a matrix, where m is the number of times v occurs in $\mathcal{W}$. Let m_t and m_{t^+} be the count of the word up to and after time t, respectively. A semantic change score is calculated for the candidate v for every $t \in \mathcal{T}$ as,

$$\overline{\boldsymbol{V}_t} = \textsc{mean}(\boldsymbol{V}_{\leq t}) \tag{7.2}$$

$$\overline{\boldsymbol{V}_{t^+}} = \textsc{mean}(\boldsymbol{V}_{>t}) \tag{7.3}$$

$$\boldsymbol{S} = \textsc{diag}(\textsc{var}(\boldsymbol{V})) \tag{7.4}$$

$$r(v, t) = (\overline{\boldsymbol{V}_t} - \overline{\boldsymbol{V}_{t^+}})^\top \boldsymbol{S}^{-1} (\overline{\boldsymbol{V}_t} - \overline{\boldsymbol{V}_{t^+}}) \tag{7.5}$$

Correction for boundaries The metric suffers from distortions at initial and final timestamps because both the mean and variance are estimated with large samples in the middle of the interval $\mathcal{T}$ and fewer samples at the boundaries. To account for this discrepancy, we add a scaling factor based on the frequency such that the metric is computed as follows.

$$r(v,t) = \sqrt{m_t}(\overline{\boldsymbol{V}_t} - \overline{\boldsymbol{V}_{t+}})^\top \boldsymbol{S}^{-1}\sqrt{m_{t+}}(\overline{\boldsymbol{V}_t} - \overline{\boldsymbol{V}_{t+}}) \tag{7.6}$$

$t^* = \mathrm{argmax}_t\, r(v,t)$ is selected as the transition point for the change. The intuition for the metric is as follows

- If a word changes in meaning at some time t, then the difference in its average embedding up to t and the average embedding after t should be high. The metric in Equation 7.6 captures this precisely by calculating the term $\overline{\boldsymbol{V}_t} - \overline{\boldsymbol{V}_{t+}}$

- Difference in average embedding can be high for seasonal or bursty changes in words (e.g. the term *turkey* is used to refer to the bird much more frequently at the time of American holidays in the winter (Shoemark et al. 2019)). To account for that we penalize the difference by their variance. If the variance is high, then the change should be given less importance and vice versa. This is also captured in the metric by the multiplicative term $\boldsymbol{S}^{-1}$.

7.2.3 Situating and aggregating usages of semantic changes

The result of the previous step is a list of semantic innovations and the time at which the meaning is deemed to have transitioned. The next step is to situate every instance of a semantic innovation as either being used in the old or new sense with respect to their time of transition. A usage is labeled as old if it is similar to other usages before the time of transition; otherwise, it is new. This is a similar task to § 5.2 but with two key differences. First, instead of assigning a continuous value score to each usage, this step only needs the usages to be categorized into one of the two labels. Second, there is no auxiliary step needed to aggregate the usages over the entire document. Owing to these differences, this step becomes decisively simpler than the model-based likelihood calculation in § 5.2.3.

To categorize every usage of a semantic innovation v, the embeddings in $\boldsymbol{V}$ are passed through a logistic regression classifier that predicts whether the usage is before the transition time or after the transition time. At the end of this step a sequence of embeddings for any semantic innovation is converted to a sequence of binary labels denoting their usage. This sequence is further filtered to keep only the usages of the word that are marked as new by the classifier.

7.2.4 Semantic leadership network induction

The previous step outputs a sequence for every semantic innovation. Recall that every usage additionally has the timestamp associated with it. This step then models these sequences as different cascades on a latent network of sources in $\mathcal{S}$ using a HP model whose intensity function is defined as follows.

$$\lambda^{(s)}(t) = \mu^{(s)} + \sum_{t_n < t} \alpha_{s_n \rightarrow s} \kappa(t - t_n), \tag{7.7}$$

where $\mu^{(s)}$ is the base intensity for source s at time t, $\alpha_{s_n \rightarrow s}$ is a pairwise influence parameter that captures the influence of source s_n on source s, and $\kappa(\cdot)$ is a time-decay kernel. In the experiments, the exponential decay kernel is $\kappa(\Delta t) = e^{-\gamma \Delta t}$ and γ is set to 1. All the parameters of the model are contrained to be non-negative and learned through stochastic gradient descent.[1]

7.3 Data

This study uses the *s2orc* dataset (Lo et al. 2020) as a starting point. Every research paper in this dataset is represented as a `JSON` object with fields such as the year of publication, body of the abstract and the main content, author names, incoming and outgoing citations, *etc*. As the focus of this study is on research papers that were published in Computational Linguistics venues, the dataset was filtered to retain only the papers that appear in the ACL

[1] Since many events can share the same timestamp, a small random offset is added to each timestamp so that every event has a unique timestamp.

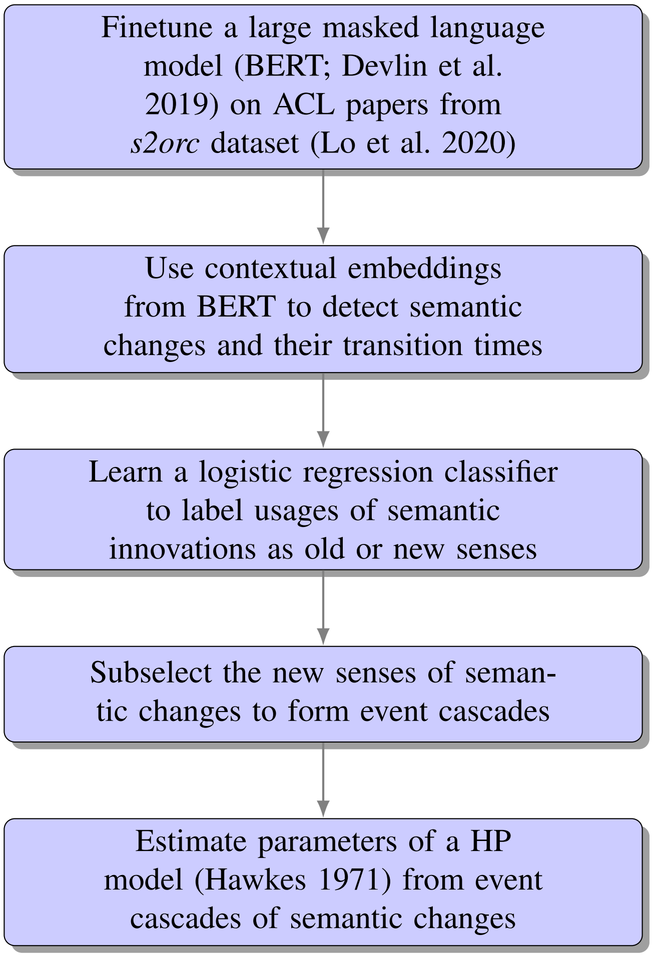

Finetune a large masked language model (BERT; Devlin et al. 2019) on ACL papers from *s2orc* dataset (Lo et al. 2020)
Use contextual embeddings from BERT to detect semantic changes and their transition times
Learn a logistic regression classifier to label usages of semantic innovations as old or new senses
Subselect the new senses of semantic changes to form event cascades
Estimate parameters of a HP model (Hawkes 1971) from event cascades of semantic changes

Figure 7.1: **Method pipeline.** Flowchart shows the complete methodological pipeline followed in this chapter.

Table 7.1: **Dataset summary.** Descriptive summary of the curated ACL corpus from *s2orc* dataset

Statistic	Value
Number of papers	32056
Years	1990–2019
Number of venues	166
Average number of citations (in-degree)	24.01
Average number of references (out-degree)	14.92
Average abstract length (number of words)	101
Average body length (number of words)	2825

anthology. This is done by checking for the presence of a valid `acl_id` field in the `JSON` object.

Though the dataset contains papers from as far back as 1965, the coverage in the early years is sparse with few or no papers in many of the early years. As a result, the data is further filtered to retain only the papers that appear from 1990 to 2019 ($\mathcal{T} = [1990, 2019]$). The *s2orc* dataset does not contain consistent information about the venue of a research paper. To resolve this, the venue of a paper is queried from the ACL anthology[2] by matching the `acl_id` from the *s2orc* dataset. For the semantic leadership network induction, venues that do not appear more than 3 times are ignored. If a paper is assigned to multiple venues, then the paper is resolved to be part of the more popular venue in the corpus.

Descriptive statistics of the curated corpus is given in Table 7.1.

7.4 Experimental setup

For this study, multilingual BERT is used as the contextualizing model. Specifically, the `bert-base-multilingual-uncased` model from the Hugging face (Wolf et al. 2019) library is used.[3] The size of the contextualized embeddings is 786 dimensions.

[2] https://www.aclweb.org/anthology/venues/

[3] https://huggingface.co/bert-base-multilingual-uncased

Finetuning Previous work has shown that the quality of the contextual embeddings improves when the pretrained BERT is finetuned on domain-specific text (e.g., Gururangan et al. 2020). In this study, the pretrained BERT model is finetuned on the abstracts of all the papers that are part of the corpus for 3 epochs to optimize the masked language modeling objective. The probability of masking is set to 15 %.

Wordpiece aggregation Since BERT learns subword embeddings by breaking tokens into wordpieces, the embeddings of the wordpieces need to be aggregated to get a representation of a token. This aggregation is done by taking the average of the wordpiece embeddings.[4]

Data preprocessing Non-English papers in the corpus are ignored from the analysis by identifying the language of the papers using `langid` (Lui et al. 2012). The vocabulary $\mathcal{V}$ is constructed by retaining words that appear at least 10 times in the abstracts and do not appear in more than 90 % abstracts. Each paper is first segmented by whitespace and then broken into chunks of 200 tokens. Only alphabetic tokens are retained. These chunks act as separate documents (see § 7.2.1) that are input to BERT.

Classifying individual usages of semantic innovations The off-the-shelf logistic regression classifier from `scikit-learn` is used to mark every individual instance of a semantic innovation as new or old. l_2 regularization is added to avoid overfitting but other than that all the inputs to the classifier are set to default. 4-fold cross-validation is performed to get the final assignment of labels from the classifier.

[4] Elementwise $\max$ as an alternative strategy of aggregation was also tried and performed similar in detecting changes

7.5 Results

7.5.1 Semantic changes

Applying the method from § 7.2.2 identifies 2092 semantic changes that capture several technical concepts in language research. The top changes, the year in which the words are deemed to change, and the contexts in which these words appear before and after the transition year are shown in Table 7.2.

The evolution of language research, from the earlier focus on syntax and sequence processing using latent variable models to the current paradigm of using deep learning, is neatly summarized by the semantic innovations that the method identifies. For example, throughout the nineties, terms such as *tokenization*, *subtree*, *connectionist*, *reflexive*, and *transducers*, among others, are recognized as semantic innovations. These terms suggest the structural approach to core NLP research during the time.

The next decade saw changes in terms such as *kernel* and *probabilistic*. These indicate some of the methodological changes that were underway during this period, with NLP research being dominated by a mix of kernel and bayesian methods during this decade (e.g., Moschitti 2004; Blei et al. 2003). Methodological innovations such as conditional random fields (Lafferty et al. 2001) and the rise of domain adaptation (e.g., Chelba et al. 2004; Daumé III 2007) is also evidenced by terms such as *conditional* and *adaptation* marked as semantic changes.

The current decade saw words such as *representations*, *deep*, and *decoder* marked to have transitioned between the years 2013 to 2015. This coincides with the time that deep learning methods started making strong inroads in language research. One prominent example of this shift is the term *attention*, shown in Figure 7.2, which shifts from its standard, broad usage to the more technical and focused usage with respect to neural networks around 2015.

Table 7.2: **Semantic change examples.** Top examples of semantic changes identified from the curated ACL corpus from the *s2orc* dataset. The relative counts are counts per million tokens. Terms such as *attention* get a new sense increasingly used later; terms such as *plan* shows semantic widening moving from strong association with dialogue to other NLP tasks; terms such as *network* and *deep* show semantic narrowing moving from disperse associations to a more narrower sense associated with neural networks.

Term	Year	Score	Relative count pre-transition	Relative count post-transition	Earlier usages	Later usages
attention	2015	2.38	126	1670	*increased **attention** over the past several years*	*parallelizable **attention** networks*
					*need to be paid **attention***	*vector of **attention** weights*
plan	2001	1.52	381	158	***plan** such a message*	***plan** recognition problems*
					*embedded in the **plan** library*	***plan** for tag generation*
network	2013	1.19	240	1000	*semantic **network** path schemata*	*deep learning **network** configurations*
					***network** of semantically related noun senses*	***network** parameters to tune*
focus	2006	0.99	451	521	*tracking local **focus***	*main **focus** of our work*
					***focus** of attention in discourse*	*the **focus** particle*
representations	2013	0.94	257	1018	*grammatical **representations***	*learning distributed **representations***
					*logical semantic **representations***	*learned **representations** across views*
deep	2014	0.94	114	417	***deep** cognitive understanding*	***deep** learning*
					***deep** syntactic features*	***deep** architectures*

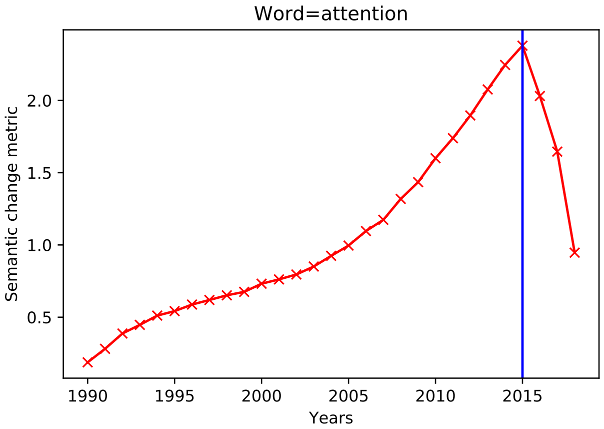


Figure 7.2: **Visual depiction of change in top example.** Semantic change in the term *attention* in *s2orc*'s ACL anthology subset. The blue line indicates the transition year for meaning change. The transition year for the term *attention* coincides with early papers that described the attention mechanism in neural networks (Bahdanau et al. 2015) that later became the bedrock of transformers architecture (Vaswani et al. 2017)

7.5.2 Labeling usages as progressive or conventional

Every usage of a semantic innovation is labeled as progressive or conventional using the contextual embeddings as features to a classifier following the setup described in § 7.2.3. The results of this classification are shown in Table 7.3. At the end of this step, the sequence of embeddings for each semantic innovation is converted to a sequence of labels with timestamps which are the event cascades used in the next step. Events that have the conventional labels are discarded in the next step.

7.5.3 Influence network induction

Recall that from the previous steps, every semantic innovation is represented as a cascade of timestamps and the publication venues. The goal of this step is to induce an influence network between the venues. For this step, the cascades are fed as input to the HP model to induce the network over the venues. The influence network between venues for semantic changes is compared to the citation network of the venues. Furthermore, the roles of the venues from the semantic leadership network are determined using the hubs and authority

Table 7.3: **Classification performance.** The performance of the classifier in distinguishing the instances of any semantic innovation as progressive or conventional. The last row shows the performance averaged across all the 2092 semantic innovations. All performance numbers are percentage values.

Word	Accuracy	Precision	Recall	F1	AUC
attention	92.0	98.8	90.7	90.2	93.5
plan	82.0	92.3	82.6	83.0	81.6
network	81.5	94.0	78.0	82.0	83.6
focus	70.8	89.2	70.8	73.0	70.9
representations	78.3	91.0	75.4	79.0	79.9
deep	84.9	90.2	82.9	85.0	85.3
average	68.7	74.9	68.8	71.8	72.1

centrality calculation, similar to chapter 6. The results from both these explorations are described next.

Exploration 1: Comparing citation network with semantic leadership network

The pagerank centrality of the two networks shown in Figure 7.3 reveal a few prominent patterns. First, popular venues such as *acl*,[5] *naacl*,[6], and *emnlp*[7] differ from each other in significant ways: *acl* and *naacl* have low centrality in the semantic leadership network but are cited much more frequently than other venues; in contrast, *emnlp* is central in both the networks. This pattern suggests that semantic innovations are more likely to emerge in *emnlp* compared to the other venues.

Second, workshops such as *bucc* and *wmt*, popular workshops in the area of machine translation, are semantic leaders even though they are not the most central when it comes to citations. This interesting pattern suggests a further line of inquiry about the direction of flow of these semantic innovations: are workshops or smaller conferences more likely to be the host of semantic innovations compared to larger conferences? This investigation could also shed some light into the prestige of these venues.

[5] https://www.aclweb.org/anthology/venues/acl/

[6] https://www.aclweb.org/anthology/venues/naacl/

[7] https://www.aclweb.org/anthology/venues/emnlp/

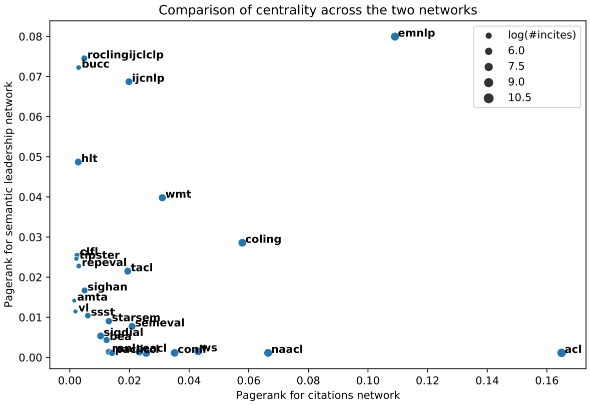


Figure 7.3: **Pagerank.** Pagerank comparison between the citation and semantic leadership network. A popular venue like *acl* has high centrality in the citations network but not in the semantic leadership network; on the other hand, workshops such as *wmt* and *bucc* are leaders on semantic changes even though they are not cited as frequently as other venues

Exploration 2: Comparing the roles of venues from the semantic leadership network

Figure 7.4 compares the different venues for the roles they play in the semantic leadership network. For this analysis, the same methodology from chapter 6 is used to distinguish between leaders and followers in the network. Once again, individual workshops such as *vl*,[8] and small-scale conferences such as *ranlp*[9] tend to be overall leaders, whereas *acl* tends to be a follower. A venue such as *semeval* tends to play a peripheral role in semantic leadership.

This exploration suggests that there is significant diversity in the roles that publication venues play in semantic leadership. That a large conference venue such as *acl* is a follower suggests that it may act as an aggregator of good innovative ideas from other venues.

7.6 Summary

This chapter shows the utility of contextual embeddings in identifying semantic changes in papers published at venues that are part of the ACL anthology. The semantic changes and

[8] https://www.aclweb.org/anthology/venues/vl/

[9] https://www.aclweb.org/anthology/venues/ranlp/

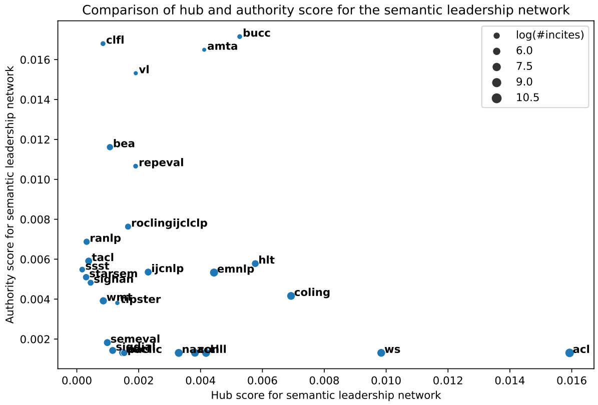


Figure 7.4: **HITS.** The hub and authority centrality of publication venues in the corpus from the semantic leadership network.

the year in which the meanings transition are identified by considering both the first and second moments of the contextual embeddings. Further, every instance of any semantic innovation is classified to mark whether it is innovative or not. The sequence of innovative instances are used to form a cascade on the latent influence network of publication venues, the weights of which can be learned using a HP model. Exploratory analysis of the semantic leadership network between the venues to the citation network suggests that the pathway of linguistic influence could be from workshops and smaller conferences such as *wmt* and *ranlp* to major conferences such as *acl*.

CHAPTER 8
CONCLUSION

8.1 Thesis summary

This thesis proposes network and text methods to model linguistic leadership in timestamped text. The individual studies propose methods to fill the gaps in existing research and in the process reveal substantive findings that span the disciplines of sociolinguistics, social science, and digital humanities.

Summary of research questions and findings

RQ1 : *Does language change spread on a behavioral network of Twitter mentions and are strong and local ties more likely to lead to adoption of language change? Can Hawkes Process models be scaled to millions of nodes to answer these questions?*

To answer this research question fully first requires establishing whether language change propagates across social ties in online networks as other sociolinguists have established in offline networks (e.g., Milroy et al. 1992; Labov 2001). Investigating the usages of non-standard terms as linguistic variables over time reveals that non-standard words spread on Twitter in the form of a contagion and whose characteristics differ based on word category: words whose origins lie in spoken communication (e.g. *hella*) follow a simple contagion but words that originate in online written communication (e.g. *lls*) follow a complex contagion (Centola et al. 2007). A more granular analysis where every word is treated as a cascade on the Twitter network reveals that linguistic influence is exerted more through densely embedded ties than the local ties in the network. Such an analysis requires modifying the Hawkes process (HP) models of network cascades to scale to millions of nodes and edges through parameter sharing, numerical approximations, and optimization (see chapter 3 and A).

RQ2 : *Under what conditions can social influence be detected from observational data about event cascades, given the presence of homophily, missing events, and misspecification?*

The analysis in RQ1 reveals a methodological concern in the parametric HP models; the generative assumptions of HP models make them a fragile candidate for the task of detecting social influence in event cascades when the assumptions are not met. To answer RQ2, first the conditions that confound the detection of social influence, such as homophily (McPherson et al. 2001), missing events in the cascade, and misspecification in the data generation process, are rigorously tested on synthetic data. Next, a new discriminative model and a non-parametric test is proposed that follows the principle of Granger causality in detecting social influence. Synthetic data analysis shows that unlike HP model, the discriminative model works well under adverse confounding conditions. The applicability of the model is shown to extend to cascades of both linguistic and non-linguistic behavior in real data (see chapter 4).

RQ3 : *Given a timestamped collection of documents, can computational methods identify which documents and producers of documents lead with respect to a semantic innovation?*

In contrast to RQ1, to answer RQ3 this thesis moves away from lexical change to lexical semantic change. First, this thesis builds upon the work of using diachronic word embeddings to detect semantic changes (Hamilton et al. 2016b), to further use these embeddings in scoring semantic changes for their *progressiveness*. For any semantic innovation, the progressiveness of its usage measures whether the specific usage is with a new or an old meaning. The proposed metric of semantic progressiveness is extended to a document which serves as a measure of a document's leadership.

Qualitative analysis shows that this measure indexes important documents in two domains: legal opinions in U.S courts and scientific abstracts in Computer Science. Further univariate and multivariate analysis shows that the proposed metric of linguistic leadership of documents correlates to other forms of influence such as the number of citations (see chapter 5) on both scientific papers and legal opinions.

Second, this thesis also proposes an extension of diachronic word embeddings to further specialize them on the producers of the documents. These *source-dependent* diachronic word embeddings are used to propose a measure of the lead (or lag) of one source over another for any semantic change. Applying this metric on semantic changes found in a collection of timestamped abolitionist newspaper articles from the nineteenth century show the leading and lagging newspapers for important concepts in the abolition discourse such as *justice* and *freedom* that were important (see chapter 6).

RQ4 : *Can lead-lag relations for individual semantic innovations be aggregated to induce a semantic leadership network on the producers of documents? Can overall leaders be identified from such a network?*

The metric of lead between the pair of newspapers on individual changes as shown in RQ3 helps in the micro-level analysis of semantic changes. RQ4, on the other hand, seeks to understand whether the semantic changes can be aggregated to find overall leaders of change, which is useful for macro-analysis. The application of the lead metric on the abolitionist newspapers corpus yields semantic leadership events which identify a pair of leader-follower newspapers for every semantic change. To account for confounds such as temporal precedence and the frequency of publishing, a statistical randomization procedure is proposed that helps retain the events from the observed data which are significant in comparison to randomized data. These individual lead-lag signals are then aggregated to form a semantic leadership network between the newspapers. The induced network reveals, on aggregate, the newpapers that introduce and pickup changes, as well as the latent pathways of influence. Centrality analysis on the induced network using pagerank (Page et al. 1999) and HITS (Kleinberg 1999) further show that newspapers edited by black and female editors (e.g. THE LILY and THE PROVINCIAL FREEMAN) play a pivotal role in shaping the abolition discourse in this period. The finding is important because it quantitatively validates the recent scholarship among historians who emphasize the once neglected role of editors from minority groups in shaping the discourse for the abolition of slavery (Sinha 2016) (see chapter 6)

RQ5 : *Can contextual word representations be used to investigate the connection between semantic leadership and other explicit markers of influence such as the citations?*

Finally, this thesis proposes a methodological pipeline based on contextualized embeddings that goes from a timestamped text corpus to inducing a semantic leadership network. The proposed method uses the embeddings of a word from a finetuned BERT model to identify the semantic changes and the time of shift in meaning. A classifier then uses the contextual embeddings as input to predict every usage as occurring before or after the shift to convert a sequence of timestamped contextual embeddings for any word into a sequence of events in the form of source and timestamp pairs. The final step in the pipeline models these event sequences as a cascade on a latent network which is recovered using HP models described in chapter 3. The application of this methodological pipeline on the propagation of technical terms in research papers published at venues from the ACL anthology provides a means to compare linguistic leadership to citation impact.

Overall thesis summary

This thesis emphasizes the use of diachronic modeling of text through computational methods to find linguistic leaders of change. In contrast to the traditional paradigm of sociolinguistics that dominantly uses apparent-time analysis, this thesis fits in the larger arc of real-time sociolinguistics. To model large text corpora from diverse sources such as social media and digital archives, this thesis proposes a number of models that scale to millions of data points (e.g., the number of nodes in a network, number of documents, number of tokens, etc). It also proposes novel metrics that can isolate semantic changes at specific times, score documents for their semantic innovativeness, and measure the semantic lead between a pair of documents sources. Finally, the thesis also demonstrates the applicability of sociolinguistic analysis in disciplines such as the sociology of science and digital humanities.

8.2 Limitations

This section describes some of the limitations in each of the studies that are integral to this thesis.

8.2.1 Lexical change

Lexical change, either in the form of new lexical items that enter the lexicon (for example, non-standard terms in chapter 3 and hashtags in chapter 4) or existing lexical items that are repurposed for new meanings (as in chapter 5, chapter 6, and chapter 7) is the focus of this thesis. The exclusive focus on word-level changes is advantageous: it is computationally easy to detect lexical changes in a large corpus compared to other forms of changes such as, syntactic of framing changes; words are a common linguistic unit across many languages; and words give both interpretability and granularity to language change analysis.

However, there are some limitations. Lexical changes can scratch the surface of a larger change that is happening. Word-level changes may prove insufficient to get the full scope of these changes. This is evident in, for example, the study of semantic change in nineteenth century abolitionist newspapers in chapter 6, where many of the words that were identified as semantic changes were, surprisingly, everyday words and did not at first glance index the ideological changes that were sweeping across the country during that time. However, closer qualitative analysis revealed that many such terms were indicative of higher-order discourse changes. This calls for a need of methods with the ability to move beyond lexical changes.

Words also do not change independently as is tacitly assumed by the computational models in this thesis; for example, the dynamics of words and the concepts they represent are shown to be correlated (e.g., Tan et al. 2017) as a result of contestation (e.g., Turney et al. 2019) or the tendency to occupy the same linguistic or social niches (e.g., Stewart et al. 2018; Ryskina et al. 2020). Without the ability to group changes, the current methods discussed in this thesis cannot distinguish the different, sometimes competing, underlying processes that result in language change.

Finally, treating words as the primary lexical unit in text is dependent on the ability of computational methods to segment text into words. While word segmentation is fairly standard in English language, for languages such as Chinese and Japanese it is not trivial (Shao et al. 2018c). This limits the direct application of using computational methods proposed in this thesis to detect changes in such languages. Though this problem can be partly alleviated by using models such as BERT in chapter 7 that have sub-words as the basic units.

8.2.2 Causation versus correlation

As noted in chapter 1, the core task of this thesis is to identify who is leading or trailing language change, which requires making causal inferences about influence: leaders of language change are innovators or early adopters of change; followers see these changes and are influenced by the language leaders in adopting change. In chapter 3 and chapter 4, I present both generative and discriminative models that can test whether sequential behavior in a network is due to influence. Yet, in general, this is an impossible task to solve on observational data alone without making key assumptions (e.g. Shalizi et al. 2011; Shalizi et al.

2016). The methods to detect influence or contagion effects in this thesis can only make inferences under *Granger causality* (Granger 1969), which states that X Granger-causes Y if prediction of Y is aided by knowledge of X. As a result, the limitations of Granger causality as a framework for causality are extended to methods in this thesis.

This question about causality is also encountered, albeit in a different form, in chapter 5, where I establish a positive relationship between semantic progressiveness, a measure of linguistic innovativeness that distinguishes some documents as leaders from others, to citation counts, a measure of impact or influence of documents, for legal opinions and scientific articles in Computer Science. However, it is impossible to say precisely and generally about whether innovativeness in language causes more citations or the direction of influence is reverse just by observational methods. The substantive findings of this thesis, even though non-trivial, should not be taken as sweeping causal claims about language change and influence.

8.2.3 Discretization

The first step in all the studies presented in this thesis — with the exception of the study in chapter 4 which assumed that the key words were provided — was to identify a set of linguistic variables that represent language change: in chapter 3 these variables were chosen by a synchronic comparison between the usage of words across metropolitian areas; in chapter 5, chapter 6, and chapter 7, the linguistic variables — words that changed in meaning — were detected automatically from data using longitudinal comparisons.

For longitudinal comparisons, particularly to detect semantic changes, all the methods discussed suffer from a common limitation: the time has to be discretized first in order to make comparisons since the models cannot inherently handle continuous timestamps.[1] This is problematic because subtle, smooth changes cannot be detected efficiently without recourse to continuous timestamps. Although contemporaneous work has proposed solu-

[1] The network cascade models discussed in chapter 3 and 4, modelled discrete events in continuous time but as mentioned before the linguistic variables were selected using synchronic comparisons

tions for this problem (e.g., Rosenfeld et al. 2018), the modeling complexity still remains a disadvantage.

8.3 Future work

This thesis offers a set of computational methods to model dynamics of language change from diachronic data. The main thrust of this thesis is in identifying linguistic leadership in change and in quantifying the importance of influence in language change. But I have shown throughout the thesis that not only is the overall topic of theoretical interest in sociolinguistics, but important in cross-cutting disciplinary areas such as computational social science and cultural analytics. I now offer some ideas on potential future research, both theoretical(§ 8.3.1) and practical applications(§ 8.3.2), that build upon the ideas and findings from this thesis.

8.3.1 Theoretical work

Language variation and change

Improving the computational methods to model language change from diachronic data is an exciting way to further the emerging field of computational sociolinguistics (Nguyen et al. 2016). With respect to this thesis, there are many improvements that could help in getting a deeper understanding of the social characteristics of language. Many existing methods fail to distinguish between different types of semantic changes, such as broadening, narrowing, pejoration, amelioration, bleaching *etc* (see discussions in Cook et al. 2010; Luo et al. 2019; Tang 2018; Dubossarsky 2018). Teasing out these differences is the first key step in unmasking the different social and linguistic processes that underlie these changes. Diachronic semantic change models that incorporate contextual embeddings may prove crucial in an overall solution to categorizing semantic changes. In a similar vein, data-driven network methods that scale to a large volume of text and network data can be used to validate and formulate new theories of language change in a network. In particular the parametric Hawkes process model that I described in § 3.4 is suitably flexible to test the effect of a range of individual (e.g., the degree of the node, node position in the network, *etc*) and dyadic characterstics (e.g., shared community, interests, affiliations *etc*) on language change. This thesis also focused primarily on the propagation of linguistic innovations. Future work should also consider incorporating the pragmatic aspects of language change such as the perceptions or attitude of the speakers towards language change (Preston 2013).

Finally, this thesis builds upon existing sociolinguistic work about the role of influence exerted across social ties that aids in the adoption of linguistic innovations. Social contagion explains the increasing popularity of innovations over time. However, many innovations do not simply grow unabated but instead undergo a subsequent sharp decline. Experimental studies on mostly non-linguistic innovations show that abandonment can be explained as a systematic divergence by a group from other social groups (Berger et al. 2007; Berger et al. 2008; Berger 2008). It has thus been theorized as a form of identity signaling from the innovators and early adopters: when innovators and early adopters from a social group see the innovation get adopted by a social group they do not identify with, they make a choice to abandon the innovation to signal their own identity. It is unclear whether this theory holds for a large population in a noisy real-world environment in comparison to well-controlled experimental settings. Moreover, it has not been rigorously put to test for linguistic innovations, which are also unique cultural markers and can be used to signal as well as evaluate group membership. A core theme in sociolinguistics is not just to point what has changed but to propose how the change has been actuated (Weinreich et al. 1968). Future work to investigate the dynamics of abandonment of innovations would fit this theme. This thesis lays the groundwork and provides methodological tools that can aid in such an investigation.

Social influence and homophily

Contagion or peer influence is a causal process on a network and is often confounded by homophily and external stimuli that produce similar behavioral patterns. In chapter 4, I presented a discriminative model that uses strong proxy features for homophily to control for its effect in cascading behavior in addition to peer influence. These proxies are high-dimensional vector representations of nodes in a network, or node embeddings, that capture the similarity between them(e.g., Perozzi et al. 2014; Grover et al. 2016; Hamilton et al. 2017). This use of node embeddings to do causal inference is still relatively unexplored but is theorized as a potential way to relax the independence assumptions that are needed in causal inference methods but fail in network applications (Yao et al. 2020). Future work could find the precise conditions under which node embeddings are effective substitutes for homophily. Node embeddings still only capture homophily that is manifested through the network structure. Future work could also incorporate embeddings from text produced by nodes as alternative or additional proxies to remove the confounding due to homophily (Keith et al. 2020).

Sociology of science

As computational methods that can ingest large amounts of data have improved, researchers have turned to applying these methods to study Science itself (Fortunato et al. 2018b). Recent data-driven research has highlighted the inequities that exist in Science (e.g., Clauset et al. 2015; Way et al. 2016; Koffi 2021). Studies have also incorporated the analysis of language, to varying extent, to show these disparities (e.g., Hofstra et al. 2020; Mohammad 2020; Wang et al. 2021). This thesis provides some methodological tools that can help future research. In particular, methods to detect linguistic innovations in text and tracking their evolution can directly link production of ideas in Science to the recognition of these ideas. Future research can look at the social and demographic attributes of the producers of these ideas and whether there are imbalances in their impact.

8.3.2 Practical applications

Modeling diffusion on social media

For many practical applications, understanding the conditions of virality on a network are important. In marketing, these can help design strategies that help products go viral (Aral et al. 2011a); for infectious diseases or transmission of misinformation, knowing the conditions of virality can, in turn, help design interventions to mitigate viral spread (e.g., Törnberg 2018). Researchers have determined structural conditions that favor a network contagion (e.g., Ugander et al. 2012; Aral et al. 2011b; Goel et al. 2016b). Yet, comparison of these conditions under a single modeling framework has not been possible. Future work should consider the use of the parametric Hawkes process (HP; § 3.4), which is suitable for such a comparison: different structural properties can be encoded as features in the parametric HP model and their effect can be tested through ablation.

The network models discussed in this thesis are not specialized to the propagation of lexical change. Diffusion of social phenomena, such as hatespeech (e.g., Ziems et al. 2020), fake news and misinformation (e.g., Shao et al. 2018a; Shao et al. 2018b), *etc* can all be modeled along the same line. However, research in this area either use powerful text analysis and basic network diffusion models that fail to account the complexity at a finer level, or surface features from text with powerful network diffusion models. Future research should address this gap going forward by integrating contemporaneous models such as BERT with the models that scale to large networks such as the ones described in chapter 3 and chapter 4.

Appendices

APPENDIX A
PARAMETRIC HAWKES PROCESS MODELING AND OPTIMIZATION

A.1 Objective Function

We estimated the parameters using constrained maximum likelihood. Given a cascade of events $\{(t_n, m_n)\}_{n=1}^{N}$, the log likelihood under our model is

$$\mathcal{L} = \sum_{n=1}^{N} \log \lambda^{(m_n)}(t_n) - \sum_{m=1}^{M} \int_0^T \lambda^{(m)}(t)\, \mathrm{d}t, \tag{A.1}$$

where T is the temporal endpoint of the cascade. Substituting in the complete definition of the per-user intensity functions from Equation 3.3 and Equation 3.4,

$$\begin{aligned}\mathcal{L} = \sum_{n=1}^{N} \log \left(\mu_{t_n}^{(m_n)} + \sum_{t_{n'} < t_n} \boldsymbol{\theta}^\top \boldsymbol{f}(m_{n'} \rightarrow m_n)\, \kappa(t_n - t_{n'}) \right) - \\ \sum_{m'=1}^{M} \int_0^T \left(\mu_t^{(m')} + \sum_{t_{n'} < t} \boldsymbol{\theta}^\top \boldsymbol{f}(m_{n'} \rightarrow m')\, \kappa(t - t_{n'}) \right) \mathrm{d}t.\end{aligned} \tag{A.2}$$

If the base intensities are constant with respect to time, then

$$\begin{aligned}\mathcal{L} = \sum_{n=1}^{N} \log \left(\mu^{(m_n)} + \sum_{t_{n'} < t_n} \boldsymbol{\theta}^\top \boldsymbol{f}(m_{n'} \rightarrow m_n)\, \kappa(t_n - t_{n'}) \right) - \\ \sum_{m'=1}^{M} \left(T\mu^{(m')} + \sum_{n=1}^{N} \boldsymbol{\theta}^\top \boldsymbol{f}(m_n \rightarrow m')\, (1 - \kappa(T - t_n)) \right),\end{aligned} \tag{A.3}$$

where the second term includes a sum over all events $n = \{1, \ldots, N\}$ that contibute to the final intensity $\lambda^{(m')}(T)$. To ease computation, however, we can rearrange the second term around the source m rather than the recipient m':

$$\mathcal{L} = \sum_{n=1}^{N} \log \left(\mu^{(m_n)} + \sum_{t_{n'} < t_n} \boldsymbol{\theta}^\top \boldsymbol{f}(m_{n'} \rightarrow m_n)\, \kappa(t_n - t_{n'}) \right) - \\ \sum_{m=1}^{M} \left(T\mu^{(m)} + \sum_{\{n : m_n = m\}} \boldsymbol{\theta}^\top \boldsymbol{f}(m \rightarrow \star)\, (1 - \kappa(T - t_n)) \right), \tag{A.4}$$

where we have introduced an aggregate feature vector $\boldsymbol{f}(m \rightarrow \star) = \sum_{m'=1}^{M} \boldsymbol{f}(m \rightarrow m')$. Because the sum $\sum_{\{n : m_n = m'\}} \boldsymbol{f}(m' \rightarrow \star)\, \kappa(T - t_n)$ does not involve either $\boldsymbol{\theta}$ or $\mu^{(1)}, \ldots, \mu^{(M)}$, we can pre-compute it. Moreover, we need to do so only for users $m \in \{1, \ldots, M\}$ for whom there is at least one event in the cascade.

A Hawkes process defined in terms of Equation 3.3 has a log likelihood that is convex in the pairwise influence parameters and the base intensities. For a parametric Hawkes process, $\alpha_{m \rightarrow m'}$ is an affine function of $\boldsymbol{\theta}$, so, by composition, the log likelihood is convex in $\boldsymbol{\theta}$ and remains convex in the base intensities.

A.2 Gradients

The first term in the log likelihood and its gradient contains a nested sum over events, which appears to be quadratic in the number of events. However, we can use the exponential decay of the kernel $\kappa(\cdot)$ to approximate this term by setting a threshold $\tau^\star$ such that $\kappa(t_n - t_{n'}) = 0$ if $t_n - t_{n'} \geq \tau^\star$. For example, if we set $\tau^\star = 24$ hours, then we approximate $\kappa(\tau^\star) = 3 \times 10^{-11} \approx 0$. This approximation makes the cost of computing the first term linear in the number of events.

The second term is linear in the number of social network connections and linear in the number of events. Again, we can use the exponential decay of the kernel $\kappa(\cdot)$ to approximate $\kappa(T - t_n) \approx 0$ for $T - t_n \geq \tau^\star$, where $\tau^\star = 24$ hours. This approximation means that we only need to consider a small number of tweets near temporal endpoint of the cascade. For each user, we also pre-computed $\sum_{\{n : m_n = m'\}} \boldsymbol{f}(m' \rightarrow \star)\, \kappa(T - t_n)$. Finally, both terms in the log likelihood and its gradient can also be trivially parallelized over users $m = \{1, \ldots, M\}$.

For a Hawkes process defined in terms of Equation 3.3, Ogata showed that additional speedups can be obtained by recursively pre-computing a set of aggregate messages for each dyad (m, m'). Each message represents the events from user m that may influence user m' at the time $t_i^{(m')}$ of their i^{th} event (Ogata 1981):

$$R_{m \rightarrow m'}^{(i)} \\ = \begin{cases} \kappa(t_i^{(m')} - t_{i-1}^{(m')})\, R_{m \rightarrow m'}^{(i-1)} + \sum_{t_{i-1}^{(m')} \leq t_j^{(m)} \leq t_i^{(m')}} \kappa(t_i^{(m')} - t_j^{(m)}) & m \neq m' \\ \kappa(t_i^{(m')} - t_{i-1}^{(m')}) \times (1 + R_{m \rightarrow m'}^{(i-1)}) & m = m'. \end{cases}$$

These aggregate messages do not involve the feature weights $\boldsymbol{\theta}$ or the base intensities, so they can be pre-computed and reused throughout parameter estimation.

For a parametric Hawkes process, it is not necessary to compute a set of aggregate messages for each dyad. It is sufficient to compute a set of aggregate messages for each possible configuration of the features. In our setting, there are only four binary features, and some combinations of features are impossible.

Because the words described in Table 3.1 are relatively rare, most of the users in our data set never used them. However, it is important to include these users in the model. Because they did not adopt these words, despite being exposed to them by users who did, their presence exerts a negative gradient on the feature weights. Moreover, such users impose a minimal cost on parameter estimation because they need to be considered only when pre-computing feature counts.

A.3 Coordinate Ascent

We optimized the log likelihood with respect to the feature weights $\boldsymbol{\theta}$ and the base intensities. Because the log likelihood decomposes over users, each base intensity $\mu^{(m)}$ is coupled with only the feature weights and not with the other base intensities. Jointly estimating all parameters is inefficient because it does not exploit this structure. We therefore used a coordinate ascent procedure, alternating between updating $\boldsymbol{\theta}$ and the base intensities. As explained in § 3.4, both $\boldsymbol{\theta}$ and the base intensities must be non-negative to ensure that intensity functions are also non-negative. At each stage of the coordinate ascent, we performed constrained optimization using the active set method of MATLAB's `fmincon` function.

APPENDIX B

MODELING DETAILS AND ROBUSTNESS CHECKS FOR SEMANTIC PROGRESSIVENESS

B.1 Learning word embeddings using Noise Contrastive Estimation

Suppose that the observed data is augmented with a set of "noise" examples $\{(\tilde{w}, w_t)\}$, where each $\tilde{w}$ is sampled from a unigram noise distribution P_n. Further assume that there are k noise examples for every real example. An alternative prediction task is to decide whether each example is from the real data $(D = 1)$ or from the noise $(D = 0)$. The cross entropy for this task is,

$$
\begin{aligned}
J = & \sum_t \log \Pr(D = 1 \mid w_t, w_{t'}) \\
& + \sum_{j=1}^{k} \log \Pr(D = 0 \mid w_t, \tilde{w}^{(j)}),
\end{aligned}
\tag{B.1}
$$

where each $\tilde{w}^{(j)}$ is drawn from P_n.

Now let us define the probability,

$$
\Pr(D = 1 \mid w_t, w_{t'}) = \frac{P(w_{t'} \mid D = 1, w_t) \Pr(D = 1)}{P(w_{t'} \mid D = 1, w_t) \Pr(D = 1) + P(w_{t'} \mid D = 0) \Pr(D = 0)}
\tag{B.2}
$$

$$
= \frac{P(w_{t'} \mid w_t)}{P(w_{t'} \mid w_t) + kP_n(w_{t'})}
\tag{B.3}
$$

$$
= \left(1 + k\frac{P_n(w_{t'})}{P(w_{t'} \mid w_t)}\right)^{-1}
\tag{B.4}
$$

$$
= \sigma\left(\mathbf{v}_{w_{t'}} \cdot \mathbf{u}_{w_t} - Z(w_t) - \log(kP_n(w_{t'}))\right)
\tag{B.5}
$$

$$
\approx \sigma\left(\mathbf{v}_{w_{t'}} \cdot \mathbf{u}_{w_t} - \log(kP_n(w_{t'}))\right),
\tag{B.6}
$$

where σ indicates the sigmoid function $\sigma(x) = (1 + \exp(-x))^{-1}$. The log-normalization term $Z(w_t) = \log \sum_{w'} \exp \mathbf{v}_{w'} \cdot \mathbf{u}_{w_t}$ can be dropped in Equation B.6 because the NCE objective is approximately "self-normalizing" when P_n has positive support over all $w \in \mathcal{V}$ (Mnih et al. 2013). We then maximize Equation B.1 by gradient ascent, which yields embeddings that are asymptotically equivalent to the optimizers of Equation 6.1 (Gutmann et al. 2010). Noise-contrastive estimation is closely related to the negative sampling objective typically employed in skipgram word embeddings, but of the two, only NCE-based embeddings can be interpreted probabilistically (Dyer 2014), as required by our approach.

B.2 Experimental settings for identifying semantic innovations

This section describes the steps taken to create a list of semantic innovations in these datasets. These innovations are then used to score every document for its progressiveness.

B.2.1 Preprocessing

For the legal documents, we stripped out HTML and used only the text. The scientific abstracts were available in plain text, but required filtering to identify English-language documents, which we performed using *langid.py* (Lui et al. 2012). In both collections, we converted the text to lowercase before proceeding, and employed *spaCy* for tokenization.[1]

B.2.2 Estimating Word Embeddings

For both document collections, the first (oldest) 500,000 documents were used to learn the early embeddings (matrices $\mathbf{V}^{\text{(old)}}$ and $\mathbf{U}^{\text{(old)}}$); the most recent 500,000 documents were used to learn the later embeddings (matrices $\mathbf{V}^{\text{(new)}}$ and $\mathbf{U}^{\text{(new)}}$). Embeddings were estimated using a public tensorflow implementation.[2] We ignored tokens with frequency below a predetermined threshold: 5 for the abstracts and 10 for the larger dataset of legal

[1] `https://spacy.io/`

[2] `https://www.tensorflow.org/tutorials/representation/word2vec`, accessed May 2019.

opinions. The maximum size of the context window was set to 10 tokens. The number of negative samples was set to 100. The NCE objective was optimized for 50 epochs and the size of the embeddings for each word was set to $d = 300$ dimensions. While most of the hyperparameters were set to the default values, the size of the embeddings was selected by evaluating on word similarity benchmarks (Luong et al. 2013).

B.2.3 Postprocessing

The number of nearest neighbors used for the computation of the metric to detect semantic changes was set to 50.

Names. In the case of legal opinions, names (e.g., of plaintiffs, defendants, and judges) pose a real difficulty in identifying genuine candidates of semantic innovations. Although names can be part of semantic innovations (e.g. *Nash equilibrium* or *Miranda rights*), names often change their distributional statistics due to real-world events rather than semantic change. To overcome this problem, we use two heuristics. We first label a small set of terms if they are names of people, organizations or places, and train a feed-forward neural network to map the embeddings of each word to the label. This method identifies terms that are distributionally similar to terms that are labeled as names. Second, we tag a randomly-selected 10% of the documents for their part of speech and obtain a distribution over parts-of-speech for each vocabulary item, using the pre-trained tagger provided by *spaCy*.[3] If a term is either (a) labelled as a name using the first heuristic or, (b) tagged as a proper noun more than 90% of the time, then it is likely to be a name and is therefore discarded from the candidates of semantic innovations.

Abbreviations. In the dataset of scientific abstracts, the mention of names is rare, but abbreviations pose a similar challenge. We identify abbreviations using a similar heuristic

[3] We used *spaCy* version 2.0.16 from `https://spacy.io/api/tagger`, accessed May 2019. The tagger was trained on the OntoNotes 5 component of the Penn Treebank.

procedure as described above: a term was judged as a likely abbreviation if it was used in all capital (majuscule) letters at least 90% of the time. However, as abbreviations can transition to the status of more typical words (e.g., *laser*), we chose to discard only those abbreviations which appear fewer than 25 times in both the early and the later set of abstracts. The abbreviations are common in the scientific abstracts and tend to be dominant as the top ranked semantic changes. For this reason, we kept a higher frequency threshold of 25 for them to balance between meaningful and spurious changes.

After applying all the steps mentioned above, we inspected the top words for both legal opinions and computer science abstracts and manually removed names and abbreviations that were not caught by these heuristics, as well as tokenization errors. For each dataset, we retain a list of the 1000 terms that underwent the most substantial semantic changes, as measured by overlap in their semantic neighborhood (described above). Words outside this list have similar embeddings over time; as a result, they are unlikely to yield large progressiveness scores for any documents, and will therefore not impact the overall results. As a robustness check, we also performed the regressions using the unfiltered list, and this did not qualitatively change the regression results.

B.3 Robustness Checks

We conducted a series of stability and robustness checks to verify that our proposed method is reliable. Learning word embeddings using NCE or similar such methods is prone to stability issues, in particular due to random initialization (Antoniak et al. 2018; Burdick et al. 2018). We ran our pipeline of learning word embeddings, identifying semantic innovations, and measuring the semantic progressiveness of documents for different random initialization.

Table B.1: **Accuracy** (in %) of word embeddings on the analogy testset (Mikolov et al. 2013b)

Runs	CourtListener		DBLP	
	Early	Later	Early	Later
1	20.5	22.7	11.8	16.6
2	20.9	22.7	11.5	16.5
3	20.9	22.4	11.8	16.6

Table B.2: **Spearman correlation** of word embeddings on the word similarity testset (Bruni et al. 2012)

Runs	CourtListener		DBLP	
	Early	Later	Early	Later
1	0.42	0.42	0.35	0.45
2	0.42	0.43	0.34	0.45
3	0.42	0.43	0.34	0.44

B.3.1 Word embeddings stability

Since our proposed calculation of progressiveness of every document relies heavily on the word embeddings, high variance in the word embeddings due to random initialization can potentially affect the calculation. We tested the performance of word embeddings under different initialization on benchmark testsets to verify that our method is quite stable. Specifically, we evaluate the quality of the word embeddings on analogy and word similarity tasks for three runs, each differing in the initialization point. The results are in Table B.1 and Table B.2 respectively.

B.3.2 Semantic innovations stability

Even though the performance on extrinsic benchmarks points to word embeddings being of similar quality irrespective of random initialization, it does not necessarily mean that there is low variance in uncovering semantic changes. To make this explicit, we show the top 10 top semantic changes identified for each run on the CourtListener text collection

Table B.3: **Top semantic changes** across different runs for Courtlistener text collection. Six changes appear in the top ten across all three runs, as shown in bold.

Runs	Top Semantic Innovations
1	***underpinned***, ***lodgment***, *recomissioned*, *disentangling*, ***entrenchment***, *forensically*, *replications*, ***fringe***, ***bonded***, ***clout***
2	***entrenchment***, *cloaks*, ***underpinned***, *replications*, *unshackled*, ***lodgment***, *origination*, ***clout***, ***bonded***,***fringe***
3	***underpinned***, ***lodgment***, ***entrenchment***, *cloaks*, *forensically*, *origination*, ***clout***, *telegraphing*,***fringe***, ***bonded***

Table B.4: **Top semantic changes** across different runs for DBLP text collection. Nine changes appear in the top ten across all three runs, as shown in bold.

Runs	Top Semantic Innovations
1	***osn***, ***ux***, ***asd***, ***ros***, ***ble***, ***mtc***, ***hesitant***, *apps*, ***nfc***, ***app***
2	***ux***, ***ble***, ***osn***, ***asd***, ***app***, ***hesitant***, ***mtc***, *ppi*, ***nfc***, ***ros***
3	***osn***, ***ux***, ***ros***, ***hesitant***, ***ble***, *ppi*, ***asd***, ***app***, ***mtc***, ***nfc***

in Table B.3 and for the DBLP collection in Table B.4.

B.3.3 Robust semantic progressiveness

The word embeddings and the discovered semantic innovations are stable despite differences in initialization. But the embeddings and the semantic innovations are dependent variables in our calculation of the semantic progressiveness. We performed another quantitative check to show that the progressiveness scores of documents are correlated across different runs. Table B.5 shows the spearman rank correlation across random pairs of runs for both the text collections. As can be seen the spearman rank correlation is extremely high, meaning that even the small amount of noise that is added to the embeddings due to initialization is canceled through the calculation of progressiveness scores.

Table B.5: **Spearman rank correlation** across random pairs of runs for both the text collections.

Runs	Scientific abstracts	Court opinions
1-2	0.994	0.995
2-3	0.996	0.993
1-3	0.992	0.995

B.4 Alternative measurement of semantic progressiveness

The results from the multivariate regressions using an alternative measure of semantic progressiveness. In this scoring scheme the progressiveness per document is calculated as the number of innovations in the document for which the progressiveness score is greater than the median progressiveness score across all semantic innovations. Table B.6 contains the results for the DBLP collection and the Table B.7 contains the results for the collection of court opinions.

Table B.6: **Poisson regression analysis of citations to scientific abstracts.** Each column indicates a model, each row indicates a predictor, and each cell contains the coefficient and, in parentheses, its standard error. Log likelihood is in millions of nats.

	M1	M2	M3
Constant	2.078	2.011	2.008
Outdegree	0.010	0.010	0.010
# Authors	0.024	0.024	0.024
Age	0.074	0.077	0.076
Length	0.002	0.002	0.002
BoWs	0.000	0.000	0.000
Prog.		0.049	
Prog. Q2			0.105
Prog. Q3			0.045
Prog. Q4			0.137
Log Lik.	-12.923	**-12.891**	-12.912

Table B.7: **Poisson regression analysis of citations to legal documents.** Each column indicates a model, each row indicates a predictor, and each cell contains the coefficient and, in parentheses, its standard error.

	M1	M2	M3
Constant	1.612	1.515	0.963
Outdegree	0.019	0.020	0.019
Age	0.011	0.012	0.017
Length	0.000	0.000	0.000
BoWs	0.000	0.000	0.000
Prog.		0.051	
Prog. Q2			0.577
Prog. Q3			0.615
Prog. Q4			0.745
Log Lik.	-429096	-427724	**-423474**

REFERENCES

[1] Lada A Adamic and Eytan Adar. “Friends and neighbors on the web”. In: *Social networks* 25.3 (2003), pp. 211–230.

[2] Eytan Adar and Lada A Adamic. “Tracking information epidemics in blogspace”. In: *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI’05)*. IEEE. 2005, pp. 207–214.

[3] Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. “Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors”. In: *Sixth International AAAI Conference on Weblogs and Social Media*. 2012.

[4] H Samy Alim. “Hip hop nation language”. In: *Linguistic anthropology: A reader* (2009), pp. 272–289.

[5] Aris Anagnostopoulos, Ravi Kumar, and Mohammad Mahdian. “Influence and correlation in social networks”. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2008, pp. 7–15.

[6] Stephen R Anderson. “Morphological change”. In: *The Routledge handbook of historical linguistics* (2015), pp. 264–285.

[7] Maria Antoniak and David Mimno. “Evaluating the stability of embedding-based word similarities”. In: *Transactions of the Association for Computational Linguistics* 6 (2018), pp. 107–119.

[8] Sinan Aral and Dylan Walker. “Creating social contagion through viral product design: A randomized trial of peer influence in networks”. In: *Management science* 57.9 (2011), pp. 1623–1639.

[9] Sinan Aral, Lev Muchnik, and Arun Sundararajan. “Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks”. In: *Proceedings of the National Academy of Sciences* 106.51 (2009), pp. 21544–21549.

[10] Sinan Aral and Dylan Walker. “Identifying influential and susceptible members of social networks”. In: *Science* 337.6092 (2012), pp. 337–341.

[11] Sinan Aral and Marshall Van Alstyne. “The diversity-bandwidth trade-off”. In: *American journal of sociology* 117.1 (2011), pp. 90–171.

[12] Ehsaneddin Asgari, Christoph Ringlstetter, and Hinrich Schütze. "Unsupervised Embedding-based Detection of Lexical Semantic Changes". In: *arXiv preprint arXiv:2005.07979* (2020).

[13] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *3rd International Conference on Learning Representations, ICLR 2015*. 2015.

[14] Guy Bailey. "Real and Apparent Time". In: *The Handbook of Language Variation and Change*. 2002, pp. 312–333.

[15] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. "The role of social networks in information diffusion". In: *Proceedings of the 21st international conference on World Wide Web*. 2012, pp. 519–528.

[16] Robert Bamler and Stephan Mandt. "Dynamic Word Embeddings". In: *International Conference on Machine Learning*. 2017, pp. 380–389.

[17] David Bamman, Chris Dyer, and Noah A Smith. "Distributed representations of geographically situated language". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2014, pp. 828–834.

[18] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. "Dont count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2014, pp. 238–247.

[19] Alexander TJ Barron, Jenny Huang, Rebecca L Spang, and Simon DeDeo. "Individuals, institutions, and innovation in the debates of the French Revolution". In: *Proceedings of the National Academy of Sciences* 115.18 (2018), pp. 4607–4612.

[20] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. "Speeded-up robust features (SURF)". In: *Computer vision and image understanding* 110.3 (2008), pp. 346–359.

[21] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[22] Jonah Berger. "Identity signaling, social influence, and social contagion". In: *Understanding peer influence in children and adolescents* (2008), pp. 181–199.

[23] Jonah Berger and Chip Heath. “Where consumers diverge from others: Identity signaling and product domains”. In: *Journal of Consumer Research* 34.2 (2007), pp. 121–134.

[24] Jonah Berger and Chip Heath. “Who drives divergence? Identity signaling, outgroup dissimilarity, and the abandonment of cultural tastes.” In: *Journal of personality and social psychology* 95.3 (2008), p. 593.

[25] Pavel Berkhin. “A Survey on PageRank Computing”. In: *Internet Mathematics* 2.1 (2005), pp. 73–120.

[26] David M Blei. “Probabilistic topic models”. In: *Communications of the ACM* 55.4 (2012), pp. 77–84.

[27] David M Blei and John D Lafferty. “Dynamic topic models”. In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 113–120.

[28] David M Blei, Andrew Y Ng, and Michael I Jordan. “Latent dirichlet allocation”. In: *Journal of Machine Learning Research* 3.Jan (2003), pp. 993–1022.

[29] Katherine Bode. *A world of fiction: digital collections and the future of literary history*. University of Michigan Press, 2018. ISBN: 9780472130856.

[30] Kees de Bot and Saskia Stoessel. “Introduction: Language change and social networks”. In: (2002).

[31] Laurel J Brinton and Elizabeth Closs Traugott. *Lexicalization and language change*. Cambridge University Press, 2005.

[32] Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. “Distributional semantics in technicolor”. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2012, pp. 136–145.

[33] Mary Bucholtz. “Why be normal?: Language and identity practices in a community of nerd girls”. In: *Language in society* 28.2 (1999), pp. 203–223.

[34] Mary Bucholtz, Nancy Bermudez, Victor Fung, Lisa Edwards, and Rosalva Vargas. “Hella Nor Cal or totally So Cal? the perceptual dialectology of California”. In: *Journal of English Linguistics* 35.4 (2007), pp. 325–352.

[35] Mary Bucholtz and Kira Hall. “Identity and interaction: A sociocultural linguistic approach”. In: *Discourse studies* 7.4-5 (2005), pp. 585–614.

[36] Mary Helen Bucholtz. “Borrowed blackness: African-American Vernacular English and European-American youth identities.” In: (1999).

[37] Laura Burdick, Jonathan K Kummerfeld, and Rada Mihalcea. “Factors Influencing the Surprising Instability of Word Embeddings”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2018, pp. 2092–2102.

[38] Joan Bybee. *Language change*. Cambridge University Press, 2015.

[39] Joan Bybee et al. *Frequency of use and the organization of language*. Oxford University Press on Demand, 2007.

[40] A Colin Cameron and Pravin K Trivedi. “Regression-based tests for overdispersion in the Poisson model”. In: *Journal of Econometrics* 46.3 (1990), pp. 347–364.

[41] Jim Casey. *Editing Forms: The Emergence of Editorship in Nineteenth-Century U.S. Periodicals*. PhD Dissertation. University of Delaware, 2017.

[42] Jim Casey. “Parsing Special Characters: Mary Ann Shadd and the Limits of Search Algorithms”. In: *Infrastructures of African American Print*. University of Wisconsin Press, 2019.

[43] Damon Centola and Michael Macy. “Complex contagions and the weakness of long ties”. In: *American journal of Sociology* 113.3 (2007), pp. 702–734.

[44] Ciprian Chelba and Alex Acero. “Adaptation of Maximum Entropy Capitalizer: Little Data Can Help a Lo”. In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. 2004, pp. 285–292.

[45] Justin Cheng, Lada A Adamic, Jon M Kleinberg, and Jure Leskovec. “Do cascades recur?” In: *Proceedings of the 25th international conference on world wide web*. 2016, pp. 671–681.

[46] David D Clarke and Brigitte Nerlich. “Word-waves: A computational model of lexical semantic change.” In: *Language & Communication* (1991).

[47] Aaron Clauset, Samuel Arbesman, and Daniel B Larremore. “Systematic inequality and hierarchy in faculty hiring networks”. In: *Science advances* 1.1 (2015), e1400005.

[48] Paul Cook and Suzanne Stevenson. “Automatically Identifying Changes in the Semantic Orientation of Words.” In: *LREC*. 2010.

[49] Ryan Cordell, David A. Smith, Abby Mullen, and Jonathan D. Fitzgerald. *Going the Rounds Virality in Nineteenth-Century American Newspapers*. University of Minnesota Press, Forthcoming.

[50] Diana Crane. "Diffusion models and fashion: a reassessment". In: *The Annals of the American Academy of Political and Social Science* 566.1 (1999), pp. 13–24.

[51] Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer, 2003.

[52] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. "Echoes of power: Language effects and power differences in social interaction". In: *Proceedings of the 21st international conference on World Wide Web*. 2012, pp. 699–708.

[53] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. "No country for old members: User lifecycle and linguistic change in online communities". In: *Proceedings of the 22nd international conference on World Wide Web*. 2013, pp. 307–318.

[54] Clare J Dannenberg. "Sociolinguistics in real time". In: *American Speech* 75.3 (2000), pp. 254–257.

[55] Alexandra D'Arcy and Sali A Tagliamonte. "Not always variable: Probing the vernacular grammar". In: *Language Variation and Change* 27.3 (2015), p. 255.

[56] Hal Daumé III. "Frustratingly Easy Domain Adaptation". In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. 2007, pp. 256–263.

[57] Ferdinand De Saussure. "Nature of the linguistic sign". In: (1916).

[58] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. "Indexing by latent semantic analysis". In: *Journal of the American society for information science* 41.6 (1990), pp. 391–407.

[59] Morris H. DeGroot and Mark J. Schervish. *Probability and statistics*. 4th. Pearson Education, 2011.

[60] Marco Del Tredici and Raquel Fernández. "Semantic Variation in Online Communities of Practice". In: *IWCS 2017-12th International Conference on Computational Semantics-Long papers*. 2017.

[61] Marco Del Tredici, Raquel Fernández, and Gemma Boleda. "Short-Term Meaning Shift: A Distributional Exploration". In: *Proceedings of the 2019 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019, pp. 2069–2075.

[62] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, pp. 4171–4186.

[63] Robin Dodsworth. "Bipartite network structures and individual differences in sound change". In: *Glossa: a journal of general linguistics* 4.1 (2019).

[64] Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. "Recurrent marked temporal point processes: Embedding event history to vector". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1555–1564.

[65] Haim Dubossarsky. "Semantic change at large: A computational approach for semantic change research". PhD thesis. Ph. D. thesis, Hebrew University of Jerusalem, Edmond and Lily Safra Center, 2018.

[66] Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. "Outta control: Laws of semantic change and inherent biases in word representation models". In: *Proceedings of the 2017 conference on empirical methods in natural language processing*. 2017, pp. 1136–1145.

[67] Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, Dominik Schlechtweg, et al. "Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change". In: *The 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019) Proceedings of the Conference*. ACL. 2019.

[68] Chris Dyer. "Notes on noise contrastive estimation and negative sampling". In: *arXiv preprint arXiv:1410.8251* (2014).

[69] Penelope Eckert. "The whole woman: Sex and gender differences in variation". In: *Language variation and change* 1.3 (1989), pp. 245–267.

[70] Penelope Eckert and Sally McConnell-Ginet. *Language and gender*. Cambridge University Press, 2013.

[71] Eugene Edgington and Patrick Onghena. *Randomization tests*. CRC Press, 2007.

[72] Jacob Eisenstein. "Identifying regional dialects in on-line social media". In: *The handbook of dialectology* (2017), pp. 368–383.

[73] Jacob Eisenstein. “Systematic patterning in phonologically-motivated orthographic variation”. In: *Journal of Sociolinguistics* 19.2 (2015), pp. 161–188.

[74] Jacob Eisenstein. “Written dialect variation in online social media”. In: *Charles Boberg, John Nerbonne, and Dom Watt, editors, Handbook of Dialectology. Wiley* (2015).

[75] Jacob Eisenstein, Brendan O’Connor, Noah A Smith, and Eric P Xing. “Diffusion of lexical change in social media”. In: *PloS one* 9.11 (2014), e113114.

[76] Jacob Eisenstein, Amr Ahmed, and Eric P Xing. “Sparse additive generative models of text”. In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. Omnipress. 2011, pp. 1041–1048.

[77] John Ernest. *Liberation historiography: African American writers and the challenge of history, 1794-1861*. University of North Carolina Press, 2004. ISBN: 9780807828533.

[78] Benjamin Fagan. *The black newspaper and the chosen nation*. The University of Georgia Press, 2016. ISBN: 9780820349404.

[79] Zsuzsanna Fagyal, Samarth Swarup, Anna María Escobar, Les Gasser, and Kiran Lakkaraju. “Centers and peripheries: Network roles in language change”. In: *Lingua* 120.8 (2010), pp. 2061–2079.

[80] John R Firth. “A synopsis of linguistic theory, 1930-1955”. In: *Studies in linguistic analysis* (1957).

[81] P. Gabrielle Foreman. *Activist sentiments: reading Black women in the nineteenth century*. The new Black studies series. University of Illinois Press, 2009. ISBN: 9780252034749.

[82] Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert-László Barabási. “Science of science”. In: *Science* 359.6379 (Mar. 2018), eaao0185+.

[83] Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. “Science of science”. In: *Science* 359.6379 (2018).

[84] F. S. Foster. “A Narrative of the Interesting Origins and (Somewhat) Surprising Developments of African-American Print Culture”. In: *American Literary History* 17.4 (2005), 714740.

[85] James H Fowler, Timothy R Johnson, James F Spriggs, Sangick Jeon, and Paul J Wahlbeck. "Network analysis and the law: Measuring the legal importance of precedents at the US Supreme Court". In: *Political Analysis* 15.3 (2007), pp. 324–346.

[86] Lea Frermann and Mirella Lapata. "A bayesian model of diachronic meaning change". In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 31–45.

[87] Thomas MJ Fruchterman and Edward M Reingold. "Graph drawing by force-directed placement". In: *Software: Practice and experience* 21.11 (1991), pp. 1129–1164.

[88] Marisa J. Fuentes. *Dispossessed lives: enslaved women, violence, and the archive*. 1st edition. Early American studies. PENN University of Pennsylvania Press, 2016. ISBN: 9780812248227.

[89] Eric Gardner. *Unexpected places: relocating nineteenth-century african american literature.* Univ. Press of Mississippi, 2011. ISBN: 9781617032110.

[90] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. "Word embeddings quantify 100 years of gender and ethnic stereotypes". In: *Proceedings of the National Academy of Sciences* 115.16 (2018), E3635–E3644.

[91] Michael Gavin. "Vector Semantics, William Empson, and the Study of Ambiguity". In: *Critical Inquiry* 44.4 (2018), 641673.

[92] Dirk Geeraerts. *Prototypicality effects in diachronic semantics: A round-up*. LAUD, 1990.

[93] Matthew Gentzkow, Bryan Kelly, and Matt Taddy. "Text as data". In: *Journal of Economic Literature* 57.3 (2019), pp. 535–74.

[94] Aaron Gerow, Yuening Hu, Jordan Boyd-Graber, David M Blei, and James A Evans. "Measuring discursive influence across scholarship". In: *Proceedings of the National Academy of Sciences* 115.13 (2018), pp. 3308–3313.

[95] Howard Giles, Nikolas Coupland, and Justine Coupland. "Accommodation theory: Communication, context, and consequence". In: *Contexts of accommodation: Developments in applied sociolinguistics*. Ed. by Howard Giles, Justine Coupland, and Nikolas Coupland. Cambridge University Press, 1991, pp. 1–68.

[96] Nabeel Gillani and Roger Levy. "Simple dynamic word embeddings for mapping perceptions in the public sphere". In: *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*. 2019, pp. 94–99.

[97] Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. "Analysing Lexical Semantic Change with Contextualised Word Representations". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 3960–3973.

[98] Rahul Goel, Sandeep Soni, Naman Goyal, John Paparrizos, Hanna Wallach, Fernando Diaz, and Jacob Eisenstein. "The social dynamics of language change in online networks". In: *International Conference on Social Informatics*. Springer. 2016, pp. 41–57.

[99] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. "The structural virality of online diffusion". In: *Management Science* 62.1 (2016), pp. 180–196.

[100] Scott A Golder and Michael W Macy. "Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures". In: *Science* 333.6051 (2011), pp. 1878–1881.

[101] Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. "Inferring networks of diffusion and influence". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 5.4 (2012), pp. 1–37.

[102] Manuel Gomez-Rodriguez, Jure Leskovec, David Balduzzi, and Bernhard Schölkopf. "Uncovering the structure and temporal dynamics of information propagation". In: *Network Science* 2.1 (2014), pp. 26–65.

[103] Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. "Simple, Interpretable and Stable Method for Detecting Words with Usage Change across Corpora". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020, pp. 538–555.

[104] John C Gower, Garmt B Dijksterhuis, et al. *Procrustes problems*. Vol. 30. Oxford University Press on Demand, 2004.

[105] Clive WJ Granger. "Investigating causal relations by econometric models and cross-spectral methods". In: *Econometrica: journal of the Econometric Society* (1969), pp. 424–438.

[106] Mark S Granovetter. "The Strength of Weak Ties'". In: *American Journal of Sociology* 78.6 (1973), pp. 1360–1380.

[107] William H Greene. *Econometric analysis*. Pearson Education, 2003.

[108] Jack Grieve. *Regional variation in written American English*. Cambridge University Press, 2016.

[109] Jack Grieve, Andrea Nini, and Diansheng Guo. “Mapping lexical innovation on American social media”. In: *Journal of English Linguistics* 46.4 (2018), pp. 293–319.

[110] Thomas L Griffiths and Michael L Kalish. “Language evolution by iterated learning with Bayesian agents”. In: *Cognitive science* 31.3 (2007), pp. 441–480.

[111] Robert A Gross and Mary Kelley, eds. *An extensive republic: print, culture, and society in the new nation, 1790-1840*. A history of the book in America. Published in Association with the American Antiquarian Society by The University of North Carolina Press, 2010. ISBN: 9780807833391.

[112] Aditya Grover and Jure Leskovec. “node2vec: Scalable feature learning for networks”. In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2016, pp. 855–864.

[113] Kristina Gulordava and Marco Baroni. “A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus.” In: *Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics*. 2011, pp. 67–71.

[114] Fangjian Guo, Charles Blundell, Hanna Wallach, and Katherine Heller. “The bayesian echo chamber: Modeling social influence via linguistic accommodation”. In: *Artificial Intelligence and Statistics*. 2015, pp. 315–323.

[115] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. “Dont Stop Pretraining: Adapt Language Models to Domains and Tasks”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 8342–8360.

[116] Michael Gutmann and Aapo Hyvärinen. “Noise-contrastive estimation: A new estimation principle for unnormalized statistical models”. In: *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*. 2010, pp. 297–304.

[117] William L Hamilton, Jure Leskovec, and Dan Jurafsky. “Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change.” In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Vol. 2016. 2016, p. 2116.

[118] William L Hamilton, Jure Leskovec, and Dan Jurafsky. “Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1. 2016, pp. 1489–1501.

[119] William L Hamilton, Rex Ying, and Jure Leskovec. "Inductive representation learning on large graphs". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, pp. 1025–1035.

[120] Zellig S Harris. "Distributional structure". In: *Word* 10.2-3 (1954), pp. 146–162.

[121] Marc Hassenzahl and Noam Tractinsky. "User experience-a research agenda". In: *Behaviour & information technology* 25.2 (2006), pp. 91–97.

[122] Alan G Hawkes. "Spectra of some self-exciting and mutually exciting point processes". In: *Biometrika* 58.1 (1971), pp. 83–90.

[123] Johannes Hellrich and Udo Hahn. "Bad Company — Neighborhoods in Neural Embedding Spaces Considered Harmful". In: *Proceedings of the 26th International Conference on Computational Linguistics*. 2016, pp. 2785–2796.

[124] Ryan James Heuser. "Word Vectors in the Eighteenth Century." In: *Digital Humanities*. 2017.

[125] Jake M Hofman, Amit Sharma, and Duncan J Watts. "Prediction and explanation in social systems". In: *Science* 355.6324 (2017), pp. 486–488.

[126] Bas Hofstra, Vivek V Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A McFarland. "The diversity–innovation paradox in science". In: *Proceedings of the National Academy of Sciences* 117.17 (2020), pp. 9284–9291.

[127] Paul J Hooper. "On some principles of grammaticalization". In: *Approaches to grammaticalization. Ámsterdam, Philadelphia: Benjamins* (1991), pp. 17–35.

[128] Bernardo A Huberman, Daniel M Romero, and Fang Wu. "Social networks that matter: Twitter under the microscope". In: *arXiv preprint arXiv:0812.1045* (2008).

[129] Matthew O Jackson. *Social and economic networks*. Princeton university press, 2010.

[130] Stefan Jänicke, Greta Franzini, Muhammad Faisal Cheema, and Gerik Scheuermann. "On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges." In: *EuroVis (STARs)*. 2015, pp. 83–103.

[131] Adam Jatowt and Kevin Duh. "A framework for analyzing semantic change of words across time". In: *IEEE/ACM Joint Conference on Digital Libraries*. IEEE. 2014, pp. 229–238.

[132] Wenying Jiang. "The relationship between culture and language". In: *ELT journal* 54.4 (2000), pp. 328–334.

[133] Xiaorui Jiang, Xiaoping Sun, and Hai Zhuge. "Towards an effective and unbiased ranking of scientific literature through mutual reinforcement". In: *Proceedings of the 21st ACM international conference on Information and knowledge management*. 2012, pp. 714–723.

[134] Matthew L. Jockers. "Computing and Visualizing the 19th-Century Literary Genome". In: *DH*. 2012.

[135] Barbara Johnstone, Neeta Bhasin, and Denise Wittkofski. "" Dahntahn" Pittsburgh: Monophthongal/aw/and Representations of Localness in Southwestern Pennsylvania". In: *American speech* 77.2 (2002), pp. 148–166.

[136] Martha S Jones. *All Bound up Together: The Woman Question in African American Public Culture, 1830-1900.* University of North Carolina Press, 2009. ISBN: 9781469605012.

[137] David Jurgens, Tyler Finethy, James McCorriston, Yi Tian Xu, and Derek Ruths. "Geolocation prediction in twitter using social networks: A critical analysis and review of current practice". In: *Ninth international AAAI conference on web and social media*. 2015.

[138] Jens Kaiser, Dominik Schlechtweg, and Sabine Schulte im Walde. "OP-IMS@ DIACR-Ita: Back to the Roots: SGNS+ OP+ CD still rocks Semantic Change Detection". In: *arXiv preprint arXiv:2011.03258* (2020).

[139] Joshua L Kalla and David E Broockman. "Campaign contributions facilitate access to congressional officials: A randomized field experiment". In: *American Journal of Political Science* 60.3 (2016), pp. 545–558.

[140] Márton Karsai, Gerardo Iniguez, Kimmo Kaski, and János Kertész. "Complex contagion process in spreading of online innovation". In: *Journal of The Royal Society Interface* 11.101 (2014), p. 20140694.

[141] Katherine Keith, David Jensen, and Brendan O'Connor. "Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 5332–5344.

[142] Rudi Keller et al. *On language change: The invisible hand in language*. Psychology Press, 1994.

[143] David Kempe, Jon Kleinberg, and Éva Tardos. “Maximizing the spread of influence through a social network”. In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2003, pp. 137–146.

[144] Paul Kerswill and Ann Williams. “Creating a new town koine: children and language change in Milton Keynes”. In: *Language in Society* (2000), pp. 65–115.

[145] Daniel Kessler and Keith Krehbiel. “Dynamics of cosponsorship”. In: *American Political Science Review* 90.3 (1996), pp. 555–566.

[146] Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. “Temporal Analysis of Language through Neural Language Models”. In: *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. 2014, pp. 61–65.

[147] Ido Kissos and Nachum Dershowitz. “OCR error correction using character correction and feature-based word classification”. In: *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*. IEEE. 2016, pp. 198–203.

[148] Lauren F. Klein. “Dimensions of Scale: Invisible Labor, Editorial Work, and the Future of Quantitative Literary Studies”. In: *PMLA* 135.1 (2020), 2339.

[149] Lauren F. Klein, Jacob Eisenstein, and Iris Sun. “Exploratory Thematic Analysis for Digitized Archival Collections”. In: *Digital Scholarship in the Humanities* 30.suppl 1 (2015), i130i141.

[150] Jon M Kleinberg. “Authoritative sources in a hyperlinked environment”. In: *Journal of the ACM (JACM)* 46.5 (1999), pp. 604–632.

[151] Marlène Koffi. “Gendered Citations at Top Economic Journals”. In: *AEA Papers and Proceedings*. Vol. 111. 2021, pp. 60–64.

[152] Austin C Kozlowski, Matt Taddy, and James A Evans. “The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings”. In: *American Sociological Review* 84.5 (2019), pp. 905–949.

[153] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. “Statistically significant detection of linguistic change”. In: *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 2015, pp. 625–635.

[154] Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. “Diachronic word embeddings and semantic shifts: a survey”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. 2018, pp. 1384–1397.

[155] Andrey Kutuzov and Mario Giulianelli. “UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection”. In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. 2020, pp. 126–134.

[156] Timothy La Fond and Jennifer Neville. “Randomization tests for distinguishing social influence and homophily effects”. In: *Proceedings of the 19th international conference on World wide web*. ACM. 2010, pp. 601–610.

[157] William Labov. “Principles of linguistic change Volume 2: Social factors”. In: *LANGUAGE IN SOCIETY-OXFORD-* 29 (2001).

[158] William Labov. “The intersection of sex and social class in the course of linguistic change”. In: *Language variation and change* 2.2 (1990), pp. 205–254.

[159] William Labov. “The linguistic consequences of being a lame”. In: *Language in society* 2.1 (1973), pp. 81–115.

[160] William Labov. “The social motivation of a sound change”. In: *Word* 19.3 (1963), pp. 273–309.

[161] William Labov, CJN Bailey, and R Shuy. “The boundaries of words and their meanings”. In: *1973* (1973), pp. 340–73.

[162] John D Lafferty, Andrew McCallum, and Fernando CN Pereira. “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”. In: *Proceedings of the Eighteenth International Conference on Machine Learning*. 2001, pp. 282–289.

[163] Severin Laicher, Gioia Baldissin, Enrique Castañeda, Dominik Schlechtweg, and Sabine Schulte im Walde. “CL-IMS@ DIACR-Ita: Volente o Nolente: BERT does not outperform SGNS on Semantic Change Detection”. In: *arXiv preprint arXiv:2011.07247* (2020).

[164] Theodoros Lappas, Evimaria Terzi, Dimitrios Gunopulos, and Heikki Mannila. “Finding effectors in social networks”. In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2010, pp. 1059–1068.

[165] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. “Computational social science”. In: *Science* 323.5915 (2009), pp. 721–723.

[166] Kalev Leetaru. “Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space”. In: *First Monday* (2011).

[167] Winfred P Lehmann. *Historical linguistics: an introduction*. Routledge, 2013.

[168] Alessandro Lenci. "Distributional semantics in linguistic and cognitive research". In: *Italian journal of linguistics* 20.1 (2008), pp. 1–31.

[169] Thomas C. Leonard. *News for all: America's coming-of-age with the press*. Oxford University Press, 1995. ISBN: 9780195064544.

[170] Kristina Lerman, Nathan O. Hodas, and Hao Wu. "Bounded Rationality in Scholarly Knowledge Discovery". In: *CoRR* abs/1710.00269 (2017). arXiv: 1710.00269.

[171] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. "Cost-effective outbreak detection in networks". In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2007, pp. 420–429.

[172] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. "Meme-tracking and the dynamics of the news cycle". In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2009, pp. 497–506.

[173] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. "Patterns of cascading behavior in large blog graphs". In: *Proceedings of the 2007 SIAM international conference on data mining*. SIAM. 2007, pp. 551–556.

[174] Omer Levy and Yoav Goldberg. "Neural word embedding as implicit matrix factorization". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2*. MIT Press. 2014, pp. 2177–2185.

[175] Michael Ley. "The DBLP computer science bibliography: Evolution, research issues, perspectives". In: *International Symposium on String Processing and Information Retrieval*. Springer. 2002, pp. 1–10.

[176] Liangda Li, Hongbo Deng, Anlei Dong, Yi Chang, and Hongyuan Zha. "Identifying and labeling search tasks via query-based hawkes processes". In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2014, pp. 731–740.

[177] Liangda Li and Hongyuan Zha. "Learning parametric models for social infectivity in multi-dimensional hawkes processes". In: *Twenty-Eighth AAAI Conference on Artificial Intelligence*. 2014.

[178] Lucy Li and David Bamman. "Characterizing English variation across social media communities with BERT". In: *Transactions of the Association for Computational Linguistics* 9 (2021), pp. 538–556.

[179] Lu Liu, Jie Tang, Jiawei Han, Meng Jiang, and Shiqiang Yang. "Mining topic-level influence in heterogeneous networks". In: *Proceedings of the 19th ACM international conference on Information and knowledge management*. 2010, pp. 199–208.

[180] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S Weld. "S2ORC: The Semantic Scholar Open Research Corpus". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 4969–4983.

[181] Marco Lui and Timothy Baldwin. "langid. py: An off-the-shelf language identification tool". In: *Proceedings of the ACL 2012 system demonstrations*. 2012, pp. 25–30.

[182] Yiwei Luo, Dan Jurafsky, and Beth Levin. "From insanely jealous to insanely delicious: Computational models for the semantic bleaching of English intensifiers". In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. 2019, pp. 1–13.

[183] Thang Luong, Richard Socher, and Christopher Manning. "Better word representations with recursive neural networks for morphology". In: *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*. 2013, pp. 104–113.

[184] Lev Manovich. *Cultural Analytics*. MIT Press, 2020.

[185] Lev Manovich. "The Science of Culture? Social Computing, Digital Humanities and Cultural Analytics". In: *Journal ISSN* 2371 (2016), p. 4549.

[186] Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. "Discovery Team at SemEval-2020 Task 1: Context-sensitive Embeddings not Always Better Than Static for Semantic Change Detection". In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. 2020, pp. 67–73.

[187] Miller McPherson, Lynn Smith-Lovin, and James M Cook. "Birds of a feather: Homophily in social networks". In: *Annual review of sociology* 27.1 (2001), pp. 415–444.

[188] Hongyuan Mei and Jason Eisner. "The Neural Hawkes Process: A Neurally Self-Modulating Multivariate Point Process". In: *NIPS*. 2017.

[189] Robert K Merton. *The sociology of science: Theoretical and empirical investigations*. University of Chicago press, 1973.

[190] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant,

Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. “Quantitative Analysis of Culture Using Millions of Digitized Books”. In: *Science* 331.6014 (2011), pp. 176–182. eprint: `http://science.sciencemag.org/content/331/6014/176.full.pdf`.

[191] Rada Mihalcea and Dragomir Radev. *Graph-based natural language processing and information retrieval*. Cambridge university press, 2011.

[192] Rada Mihalcea and Vivi Nastase. “Word epoch disambiguation: Finding how words change over time”. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2012, pp. 259–263.

[193] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. “Distributed representations of words and phrases and their compositionality”. In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.

[194] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. “Efficient estimation of word representations in vector space”. In: *arXiv preprint arXiv:1301.3781* (2013).

[195] James Milroy and Lesley Milroy. “Linguistic change, social network and speaker innovation1”. In: *Journal of Linguistics* 21.2 (1985), pp. 339–384.

[196] Lesley Milroy. *Language and social networks*. Wiley-Blackwell, 1987.

[197] Lesley Milroy and James Milroy. “Social network and social class: Toward an integrated sociolinguistic model”. In: *Language in society* 21.1 (1992), pp. 1–26.

[198] Lesley Milroy and Carmen Llamas. “Social networks”. In: *The handbook of language variation and change* (2013), pp. 407–427.

[199] Andriy Mnih and Koray Kavukcuoglu. “Learning word embeddings efficiently with noise-contrastive estimation”. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*. 2013, pp. 2265–2273.

[200] Saif Mohammad. “Gender Gap in Natural Language Processing Research: Disparities in Authorship and Citations”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 7860–7870.

[201] Bjarke Mønsted, Piotr Sapieżyński, Emilio Ferrara, and Sune Lehmann. “Evidence of complex contagion of information in social media: An experiment using Twitter bots”. In: *PloS one* 12.9 (2017), e0184148.

[202] Franco Moretti. “Conjectures on world literature”. In: *New left review* 1 (2000), p. 54.

[203] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. “Is the Sample Good Enough? Comparing Data from Twitter’s Streaming API with Twitter’s Firehose.” In: *Proceedings of ICWSM*. 2013.

[204] Alessandro Moschitti. “A study on convolution kernels for shallow statistic parsing”. In: *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*. 2004, pp. 335–342.

[205] Mark Newman. *Networks*. Oxford university press, 2018.

[206] Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska de Jong. “Computational sociolinguistics: A survey”. In: *Computational linguistics* 42.3 (2016), pp. 537–593.

[207] Dong Nguyen, Maria Liakata, Simon DeDeo, Jacob Eisenstein, David Mimno, Rebekah Tromble, and Jane Winters. “How we do things with words: Analyzing text as social and cultural data”. In: *Frontiers in Artificial Intelligence* 3 (2020), p. 62.

[208] Dong-Phuong Nguyen. “Text as social and cultural data: a computational perspective on variation in text”. In: (2017).

[209] Eugene Nida. “Language, culture and translation”. In: *Foreign Languages Journal* 115.3 (1998), pp. 29–33.

[210] Partha Niyogi and Robert C Berwick. “A dynamical systems model for language change”. In: *Complex Systems* 11.3 (1997), pp. 161–204.

[211] Yosihiko Ogata. “On Lewis’ simulation method for point processes”. In: *IEEE transactions on information theory* 27.1 (1981), pp. 23–31.

[212] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. *The PageRank citation ranking: Bringing order to the web.* Tech. rep. Stanford InfoLab, 1999.

[213] Umashanthi Pavalanathan and Jacob Eisenstein. “Confounds and Consequences in Geotagged Twitter Data”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015, pp. 2138–2148.

[214] Jeffrey Pennington, Richard Socher, and Christopher D Manning. “Glove: Global vectors for word representation”. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[215] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. “Deepwalk: Online learning of social representations”. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2014, pp. 701–710.

[216] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. “Deep Contextualized Word Representations”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2018, pp. 2227–2237.

[217] Alexander M Petersen, Joel Tenenbaum, Shlomo Havlin, and H Eugene Stanley. “Statistical laws governing fluctuations in word use from word birth to word death”. In: *Scientific reports* 2.1 (2012), pp. 1–9.

[218] Carla L. Peterson. *Doers of the word: African-American women speakers and writers in the North (1830-1880)*. Race and American culture. Oxford University Press, 1995. ISBN: 9780195085198.

[219] Janet B Pierrehumbert. “The dynamic lexicon”. In: *Handbook of laboratory phonology* (2012), pp. 173–183.

[220] Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. “Predicting overt display of power in written dialogs”. In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2012, pp. 518–522.

[221] Dennis R Preston. “The influence of regard on language variation and change”. In: *Journal of Pragmatics* 52 (2013), pp. 93–104.

[222] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”. In: *Journal of Machine Learning Research* 21.140 (2020), pp. 1–67.

[223] Jakob Gulddahl Rasmussen. “Temporal point processes: the conditional intensity function”. In: (2011).

[224] Helena Raumolin-Brunberg. “Leaders of linguistic change in Early Modern England”. In: *Corpus-based studies of diachronic English* (2006), pp. 115–34.

[225] Jane Rhodes. *Mary Ann Shadd Cary: the Black press and protest in the nineteenth century*. Indiana University Press, 1998. ISBN: 9780253334466.

[226] Ian Roberts. *Diachronic syntax*. Oxford University Press, 2007.

[227] Alexander Robertson, Farhana Ferdousi Liza, Dong Nguyen, Barbara McGillivray, and Scott A Hale. “Semantic Journeys: Quantifying Change in Emoji Meaning from 2012-2018”. In: *arXiv preprint arXiv:2105.00846* (2021).

[228] Everett M Rogers. *Diffusion of innovations*. Simon and Schuster, 2010.

[229] Daniel M Romero, Brendan Meeder, and Jon Kleinberg. “Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter”. In: *Proceedings of the 20th international conference on World wide web*. 2011, pp. 695–704.

[230] Alex Rosenfeld and Katrin Erk. “Deep Neural Models of Semantic Shift”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Vol. 1. 2018, pp. 474–484.

[231] Maja Rudolph and David Blei. “Dynamic embeddings for language evolution”. In: *Proceedings of the 2018 World Wide Web Conference*. 2018, pp. 1003–1011.

[232] Maria Ryskina, Ella Rabinovich, Taylor Berg-Kirkpatrick, David R Mortensen, and Yulia Tsvetkov. “Where New Words Are Born: Distributional Semantic Analysis of Neologisms and Their Semantic Neighborhoods”. In: *Proceedings of the Society for Computation in Linguistics 2020*. 2020, pp. 313–322.

[233] Eyal Sagi, Stefan Kaufmann, and Brady Clark. “Tracing semantic change with latent semantic analysis”. In: *Current Methods in Historical Semantics* 73 (2011), pp. 161–183.

[234] Matthew Salganik. *Bit by bit: Social research in the digital age*. Princeton University Press, 2019.

[235] Gillian Sankoff and Hélène Blondeau. “Language change across the lifespan:/r/in Montreal French”. In: *Language* (2007), pp. 560–588.

[236] Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. “A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 732–746.

[237] Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. “Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 2018, pp. 169–174.

[238] Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. “SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection”. In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. 2020, pp. 1–23.

[239] Asaf Shabtai, Yuval Fledel, Uri Kanonov, Yuval Elovici, Shlomi Dolev, and Chanan Glezer. “Google android: A comprehensive security assessment”. In: *IEEE Security & Privacy* (2010), pp. 35–44.

[240] Mary Ann Shadd. “Adieu”. In: *Provincial Freeman* (1855).

[241] Cosma Rohilla Shalizi and Edward McFowland III. “Estimating causal peer influence in homophilous social networks by inferring latent locations”. In: *arXiv preprint arXiv:1607.06565* (2016).

[242] Cosma Rohilla Shalizi and Andrew C Thomas. “Homophily and contagion are generically confounded in observational social network studies”. In: *Sociological Methods & Research* 40.2 (2011), pp. 211–239.

[243] Chengcheng Shao, Pik-Mai Hui, Lei Wang, Xinwen Jiang, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. “Anatomy of an online misinformation network”. In: *PloS one* 13.4 (2018), e0196087.

[244] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. “The spread of low-credibility content by social bots”. In: *Nature communications* 9.1 (2018), pp. 1–9.

[245] Yan Shao, Christian Hardmeier, and Joakim Nivre. “Universal word segmentation: Implementation and interpretation”. In: *Transactions of the Association for Computational Linguistics* 6 (2018), pp. 421–435.

[246] Devyani Sharma and Robin Dodsworth. “Language Variation and Social Networks”. In: *Annual Review of Linguistics* 6 (2020), pp. 341–361.

[247] Anna Shechtman. “Command of Media’s Metaphors”. In: *Critical Inquiry* 47.2 (2020).

[248] Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. “Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, pp. 66–76.

[249] Yanchuan Sim, Brice DL Acree, Justin H Gross, and Noah A Smith. “Measuring ideological proportions in political speeches”. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 2013, pp. 91–101.

[250] William Simpson. “QAP: The quadratic assignment procedure”. In: *North American STATA users group meeting*. 2001, pp. 12–13.

[251] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-june Paul Hsu, and Kuansan Wang. "An overview of Microsoft academic service (MAS) and applications". In: *Proceedings of the 24th International Conference on World Wide Web*. ACM. 2015, pp. 243–246.

[252] M. Sinha. *The Slave's Cause: A History of Abolition*. Yale University Press, 2016. ISBN: 9780300181371.

[253] David A Smith, Ryan Cordell, and Elizabeth Maddock Dillon. "Infectious texts: Modeling text reuse in nineteenth-century newspapers". In: *2013 IEEE International Conference on Big Data*. IEEE. 2013, pp. 86–94.

[254] Noah A Smith. "Contextual word representations: putting words into computers". In: *Communications of the ACM* 63.6 (2020), pp. 66–74.

[255] Ray Smith. "An overview of the Tesseract OCR engine". In: *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*. Vol. 2. IEEE. 2007, pp. 629–633.

[256] Mark D Smucker, James Allan, and Ben Carterette. "A comparison of statistical significance tests for information retrieval evaluation". In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM. 2007, pp. 623–632.

[257] Sandeep Soni, Lauren Klein, and Jacob Eisenstein. "Abolitionist Networks: Modeling language change in nineteenth-century activist newspapers". In: *Journal of Cultural Analytics* (2021).

[258] Sandeep Soni, Lauren Klein, and Jacob Eisenstein. "Correcting whitespace errors in digitized historical texts". In: *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. 2019, pp. 98–103.

[259] Sandeep Soni, Shawn Ling Ramirez, and Jacob Eisenstein. "Detecting social influence in event Cascades by comparing discriminative rankers". In: *The 2019 ACM SIGKDD Workshop on Causal Discovery*. 2019, pp. 78–99.

[260] Sandeep Soni, Kristina Lerman, and Jacob Eisenstein. "Follow the leader: Documents on the leading edge of semantic change get more citations". In: *Journal of the Association for Information Science and Technology* 72.4 (2021), pp. 478–492.

[261] Derrick R Spires. *The Practice of Citizenship: Black Politics and Print Culture in the Early United States.* 2019. ISBN: 9780812295771.

[262] Lauren Squires. “Enregistering internet language”. In: *Language in Society* 39.4 (2010), pp. 457–492.

[263] Ian Stewart and Jacob Eisenstein. “Making fetch happen: The influence of social and linguistic context on nonstandard word growth and decline”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018, pp. 4360–4370.

[264] Ann Laura Stoler. *Along the archival grain: epistemic anxieties and colonial common sense*. Princeton University Press, 2009. ISBN: 9780691146362.

[265] Saatviga Sudhahar, Gianluca De Fazio, Roberto Franzosi, and Nello Cristianini. “Network analysis of narrative content in large corpora”. In: *Natural Language Engineering* 21.1 (2015), p. 81.

[266] Sali A Tagliamonte and Derek Denis. “Linguistic ruin? LOL! Instant messaging and teen language”. In: *American speech* 83.1 (2008), pp. 3–34.

[267] Sali A Tagliamonte and Alexandra D’Arcy. “Peaks beyond phonology: Adolescence, incrementation, and language change”. In: *Language* (2009), pp. 58–108.

[268] Nina Tahmasebi, Lars Borin, and Adam Jatowt. “Survey of Computational Approaches to Diachronic Conceptual Change”. In: *arXiv preprint arXiv:1811.06278* (2018).

[269] Chenhao Tan, Dallas Card, and Noah A Smith. “Friendships, Rivalries, and Trysts: Characterizing Relations between Ideas in Texts”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2017, pp. 773–783.

[270] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. “Line: Large-scale information network embedding”. In: *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 2015, pp. 1067–1077.

[271] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. “Arnetminer: extraction and mining of academic social networks”. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2008, pp. 990–998.

[272] Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. “Social influence analysis in large-scale networks”. In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2009, pp. 807–816.

[273] Xuri Tang. “A state-of-the-art of semantic change computation”. In: *Natural Language Engineering* 24.5 (2018), pp. 649–676.

[274] Petter Törnberg. “Echo chambers and viral misinformation: Modeling fake news as complex contagion”. In: *PloS one* 13.9 (2018), e0203958.

[275] Elizabeth Closs Traugott and Richard B Dasher. *Regularity in semantic change*. Vol. 97. Cambridge University Press, 2001.

[276] Michel-Rolph Trouillot. *Silencing the past: power and the production of history*. Beacon Press, 1995. ISBN: 9780807043103.

[277] Peter Trudgill. “Linguistic change and diffusion: description and explanation in sociolinguistic dialect geography”. In: *Language in Society* 3.2 (1974), pp. 215–246.

[278] Peter Trudgill. “Sex, covert prestige and linguistic change in the urban British English of Norwich”. In: *Language in society* 1.2 (1972), pp. 179–195.

[279] Peter Trudgill. *Sociolinguistics: An introduction to language and society*. Penguin UK, 2000.

[280] Peter Trudgill. “Vernacular universals and the sociolinguistic typology of English dialects”. In: *Vernacular Universals and Language Contacts*. Routledge, 2009, pp. 310–328.

[281] Peter D Turney and Patrick Pantel. “From frequency to meaning: Vector space models of semantics”. In: *Journal of Artificial Intelligence Research* 37 (2010), pp. 141–188.

[282] Peter D Turney and Saif M Mohammad. “The natural selection of words: Finding the features of fitness”. In: *PloS one* 14.1 (2019), e0211512.

[283] Johan Ugander, Lars Backstrom, Cameron Marlow, and Jon Kleinberg. “Structural diversity in social contagion”. In: *Proceedings of the National Academy of Sciences* 109.16 (2012), pp. 5962–5966.

[284] Ted Underwood. “A Genealogy of Distant Reading.” In: *DHQ: Digital Humanities Quarterly* 11.2 (2017).

[285] Ted Underwood, David Bamman, and Sabrina Lee. “The Transformation of Gender in English-Language Fiction”. In: *Cultural Analytics* (2018).

[286] Utkarsh Upadhyay, Abir De, and Manuel Gomez-Rodrizuez. “Deep reinforcement learning of marked temporal point processes”. In: *Proceedings of the 32nd Inter-*

national Conference on Neural Information Processing Systems. 2018, pp. 3172–3182.

[287] Nicolas Usunier, David Buffoni, and Patrick Gallinari. “Ranking with ordered weighted pairwise classification”. In: *Proceedings of the 26th annual international conference on machine learning*. ACM. 2009, pp. 1057–1064.

[288] Thomas W Valente. “Social network thresholds in the diffusion of innovations”. In: *Social networks* 18.1 (1996), pp. 69–89.

[289] Marc Van Opijneni. “Citation Analysis and Beyond: in Search of Indicators Measuring Case Law Importance”. In: *Legal Knowledge and Information Systems: JURIX 2012: the 25th Annual Conference*. Vol. 250. IOS Press. 2012, p. 95.

[290] A Van Raan. “Scientometrics: State-of-the-art”. In: *Scientometrics* 38.1 (1997), pp. 205–218.

[291] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. “Attention is All you Need”. In: *NIPS*. 2017.

[292] Victor Veitch, Yixin Wang, and David Blei. “Using embeddings to correct for unobserved confounding in networks”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 13769–13779.

[293] Rob Voigt, Nicholas P Camp, Vinodkumar Prabhakaran, William L Hamilton, Rebecca C Hetey, Camilla M Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L Eberhardt. “Language from police body camera footage shows racial disparities in officer respect”. In: *Proceedings of the National Academy of Sciences* 114.25 (2017), pp. 6521–6526.

[294] Lucy Lu Wang, Gabriel Stanovsky, Luca Weihs, and Oren Etzioni. “Gender trends in computer science authorship”. In: *Communications of the ACM* 64 (2021), pp. 78 –84.

[295] Xuerui Wang and Andrew McCallum. “Topics over time: a non-Markov continuous-time model of topical trends”. In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2006, pp. 424–433.

[296] Duncan J Watts and Peter Sheridan Dodds. “Influentials, networks, and public opinion formation”. In: *Journal of consumer research* 34.4 (2007), pp. 441–458.

[297] Samuel F Way, Daniel B Larremore, and Aaron Clauset. “Gender, productivity, and prestige in computer science faculty hiring networks”. In: *Proceedings of the 25th International Conference on World Wide Web*. 2016, pp. 1169–1179.

[298] Uriel Weinreich, William Labov, and Marvin Herzog. "Empirical foundations for a theory of language change". In: *Directions for historical linguistics* (1968), pp. 97–188.

[299] Jason Weston, Samy Bengio, and Nicolas Usunier. "Wsabie: Scaling up to large vocabulary image annotation". In: *Twenty-Second International Joint Conference on Artificial Intelligence*. 2011.

[300] Derry Tanti Wijaya and Reyyan Yeniterzi. "Understanding semantic change of words over centuries". In: *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web*. ACM. 2011, pp. 35–40.

[301] C Wilson and V Mihalicek. *Language files: Materials for an introduction to language and linguistics*. Columbus: Ohio State University Press, 2011.

[302] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. "HuggingFace's Transformers: State-of-the-art natural language processing". In: *arXiv preprint arXiv:1910.03771* (2019).

[303] Walt Wolfram. "The linguistic variable: Fact and fantasy". In: *American Speech* 66.1 (1991), pp. 22–32.

[304] Hongteng Xu, Dixin Luo, and Hongyuan Zha. "Learning Hawkes Processes from Short Doubly-Censored Event Sequences". In: *ICML*. 2017.

[305] Jaewon Yang and Jure Leskovec. "Modeling information diffusion in implicit networks". In: *Proceedings of ICDM*. 2010.

[306] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. "XLNet: Generalized Autoregressive Pretraining for Language Understanding". In: *Advances in Neural Information Processing Systems* 32 (2019), pp. 5753–5763.

[307] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. "A Survey on Causal Inference." In: (2020).

[308] Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. "Dynamic word embeddings for evolving semantic discovery". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM. 2018, pp. 673–681.

[309] Dani Yogatama, Michael Heilman, Brendan O'Connor, Chris Dyer, Bryan R. Routledge, and Noah A. Smith. "Predicting a Scientific Community's Response to an

Article". In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. 2011, pp. 594–604.

[310] Anna A Zalizniak, Maria Bulakh, Dmitrij Ganenkov, Ilya Gruntov, Timur Maisak, and Maxim Russo. "The catalogue of semantic shifts as a database for lexical semantic typology". In: *Linguistics* 50.3 (2012), pp. 633–669.

[311] Christine Zeller. "The Investigation of a Sound Change in Progress: /ae/to/e/in Midwestern American English". In: *Journal of English linguistics* 25.2 (1997), pp. 142–155.

[312] Qingyuan Zhao, Murat A Erdogdu, Hera Y He, Anand Rajaraman, and Jure Leskovec. "Seismic: A self-exciting point process model for predicting tweet popularity". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2015, pp. 1513–1522.

[313] Ke Zhou, Hongyuan Zha, and Le Song. "Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes". In: *Artificial Intelligence and Statistics*. 2013, pp. 641–649.

[314] Jian Zhu and David Jurgens. "The structure of online social networks modulates the rate of lexical change". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021, forthcoming.

[315] Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. "Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis". In: *arXiv preprint arXiv:2005.12423* (2020).

[316] Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. "Transformer hawkes process". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 11692–11702.

VITA

Sandeep Soni was born in the quiet city of Amravati in the state of Maharashtra in India. He spent his childhood in Ahmednagar with his parents and older brother. He grew up playing Cricket, solving puzzles, and collecting coins — a passion he shared with his brother. After finishing high school at 17, he moved to the city of Pune, where he went to the Pune Institute of Computer Technology (part of Pune University) to study Computer Engineering. After a fun 4 years of undergrad, he joined a research startup, Oneirix laboratories, where he learned for 2 years before moving to Atlanta to do a masters in Computer Science at Georgia Institute of Technology. Before joining the PhD program at Georgia Institute of Technology, he worked for a year at Yahoo!