Received: 1 November 2019 | Revised: 2 August 2020 | Accepted: 8 September 2020

DOI: 10.1002/asi.24421

RESEARCH ARTICLE

JASIST WILEY

# Follow the leader: Documents on the leading edge of semantic change get more citations

**Sandeep Soni**[1] | **Kristina Lerman**[2] | **Jacob Eisenstein**[3]

[1]School of Interactive Computing, Georgia Institute of Technology, Atlanta, Georgia

[2]Information Sciences Institute, University of Southern California, Los Angeles, California

[3]Google Research, Seattle, Washington

**Correspondence**
Sandeep Soni, Georgia Institute of Technology, Atlanta, USA.
Email: sandeepsoni@gatech.edu

**Funding information**
Air Force Office of Scientific Research, Grant/Award Number: FA9550-17-1-0327; Defense Advanced Research Projects Agency, Grant/Award Number: W911NF-19-2-0271; Division of Information and Intelligent Systems, Grant/Award Number: 1452443

**Abstract**

Diachronic word embeddings—vector representations of words over time—offer remarkable insights into the evolution of language and provide a tool for quantifying sociocultural change from text documents. Prior work has used such embeddings to identify shifts in the meaning of individual words. However, simply knowing that a word has changed in meaning is insufficient to identify the instances of word usage that convey the historical meaning or the newer meaning. In this study, we link diachronic word embeddings to documents, by situating those documents as leaders or laggards with respect to ongoing semantic changes. Specifically, we propose a novel method to quantify the degree of semantic progressiveness in each word usage, and then show how these usages can be aggregated to obtain scores for each document. We analyze two large collections of documents, representing legal opinions and scientific articles. Documents that are scored as semantically progressive receive a larger number of citations, indicating that they are especially influential. Our work thus provides a new technique for identifying lexical semantic leaders and demonstrates a new link between progressive use of language and influence in a citation network.

## 1 | INTRODUCTION

Languages are continuously evolving (Weinreich, Labov, & Herzog, 1968), and one of the particular salient aspects of language change is how elements such as words are repurposed to new meanings (Traugott & Dasher, 2001). Word embeddings—representations of words as vectors in high-dimensional spaces—can identify semantic changes in text documents by tracking shifts in each word's distributional neighborhood (Kutuzov, Øvrelid, Szymanski, & Velldal, 2018). However, these methods treat each word in isolation and do not indicate where change takes place, which documents or passages introduce new meanings for words, and which lag behind in adopting semantic changes?

The ability to identify documents in the vanguard of linguistic change would yield valuable insights into the life cycle of new ideas: for example, by making it possible to identify and support innovation in science (Fortunato et al., 2018), and would provide new evidence about the social processes underlying linguistic and scholarly influence (Gerow, Hu, Boyd-Graber, Blei, & Evans, 2018). As a step toward this goal, we propose a simple quantitative technique for identifying the leading examples of ongoing semantic changes. Our method builds directly on the embedding-based techniques for detecting changes in a large corpus of documents and takes the form of a likelihood-ratio comparison between "older" and "newer" embedding models. Usages that are better aligned with the newer embedding model can be

Jacob Eisenstein is an Adjunct Associate Professor at Georgia Tech and a Research Scientist at Google.

  

considered to be more semantically "progressive," in the sense of reflecting newer word meanings.

Using large data sets of legal opinions and scientific research abstracts produced over a long period of time, we demonstrate that more semantically advanced usages are indeed associated with documents that are landmarks in their respective fields, such as prominent Supreme Court rulings and foundational research papers. We further formalize these insights by demonstrating a novel relationship between semantic progressiveness and citation counts: in both domains, semantically progressive documents receive more citations, even after controlling for document content and a range of structural factors. While previous work has identified connections between word frequency and impact, we are the first to link semantic changes to citation networks. To summarize the contributions of this article:

- We identify markers of semantic change in scientific articles and legal opinions (both in English). Legal opinions have not previously been analyzed with respect to dynamic word embeddings and have received little attention in natural language processing.
- We propose a novel method to score documents on their semantic progressiveness, thereby identifying documents on the vanguard of semantic change.
- We show that documents at the vanguard of semantic change tend to be more influential in citation networks.

Diachronic word embeddings help in identifying the points of transition in the language, especially as it pertains to semantic changes. Past research in information science, especially around the science of science, has identified the importance of transitions in the scholarly process—for example, in determining the impact of scientific works (Gerow et al., 2018), understanding the impact of individual topics (Yan, 2015), studying the formation of interdisciplinary research areas (Xu et al., 2018), and in mapping out the structure of scientific communities over time (Boyack & Klavans, 2014). As a result, we see this work as relevant for a broader objective of understanding the scholarly process through the transitional aspects of language use and its link to scholarly outcomes. Our proposed approach provides a quantitative way to identify semantically innovative documents and study their dynamics. A secondary contribution to information science is to provide a new link between two orthogonal perspectives on document collections: content analysis and citation structures. While the topic modeling literature has offered models that link citations to sets of co-occurring words (e.g., Ding, 2011; Nallapati, Ahmed, Xing, & Cohen, 2008), we offer a more fine-grained lexical semantic perspective by connecting incoming citations to leadership on changes in the meanings of individual words, showcasing the potential of diachronic word embeddings as a tool for information science.[1]

## 2 | MEASURING SEMANTIC PROGRESSIVENESS

Diachronic word embeddings make it possible to measure lexical semantic change over time (e.g., Hamilton, Leskovec, & Jurafsky, 2016b; Kulkarni, Al-Rfou, Perozzi, & Skiena, 2015). In standard word embeddings, each word type is associated with a vector of real numbers, based on its distributional statistics (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Turney & Pantel, 2010). 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, our method is comprised of four steps: (a) learning diachronic embeddings of words; (b) identifying semantic innovations using their diachronic embeddings; (c) scoring each usage by its position with respect to the semantic change; and (d) aggregating these scores by document. A schematic of the entire pipeline is shown in Figure 1. We now describe each of these steps in detail.

### 2.1 | Estimating word embeddings

Several methods to learn diachronic word embeddings have been proposed (e.g., Bamler & Mandt, 2017; Frermann & Lapata, 2016; Hamilton et al., 2016b; Rosenfeld & Erk, 2018). In this work, we use the method proposed by Hamilton et al. (2016b) as it is conceptually straightforward and offers flexibility in the choice of the embedding algorithm. The core of this approach is to fit embedding models to distinct time slices of the corpora, and then align the resulting 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, $\mathcal{W} = \{(w_{i,1}, w_{i,2}, ..., w_{i,T_i})\}_{i=1}^{N}$, where $N$ is the number of documents in the corpus, $i$ indexes an individual document whose length is $T_i$, and each $w_{i,t} \in$. For each corpus, we estimate a set of word embeddings on the single vocabulary $\mathcal{V}$. Following Hamilton et al. (2016b), we estimate skipgram embeddings (Mikolov, Sutskever, et al., 2013), which are based on the objective of predicting context words $w_{t'}$ conditioned on a target word $w_t$.

Although the mathematical details of skipgram word embeddings are well known, they are crucial to our

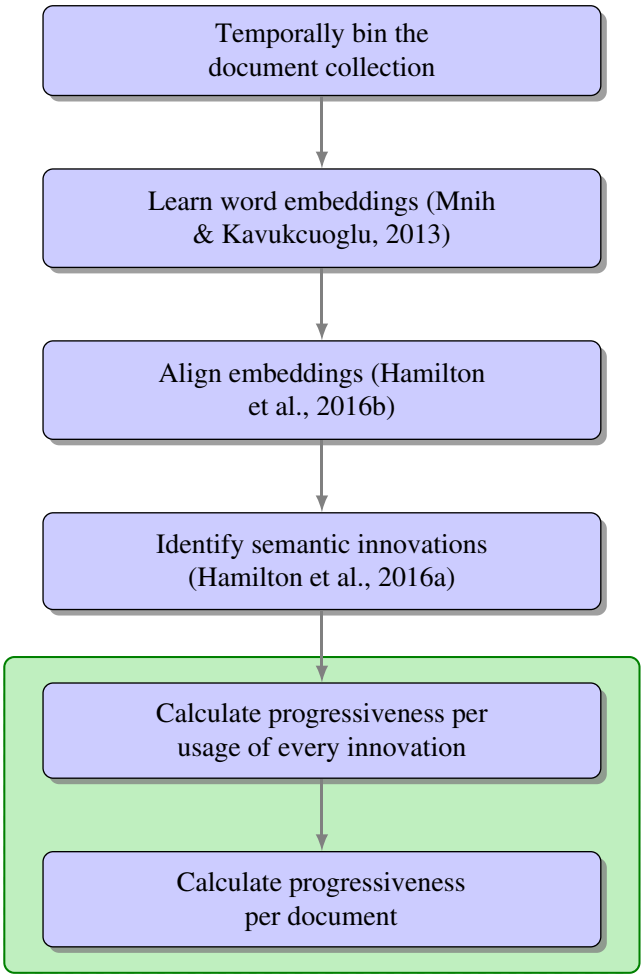


**FIGURE 1** Flowchart shows our complete pipeline and highlights (in green) our methodological contributions [Color figure can be viewed at wileyonlinelibrary.com]

method for situating individual usages of words with respect to ongoing semantic changes. For this reason, we present a brief review. Omitting the document index $i$, the skipgram objective is based on the probability,

$$P(w_{t'}|w_t) \propto \exp\left(\mathbf{v}_{w_{t'}} \cdot \mathbf{u}_{w_t}\right), \tag{1}$$

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 (1) requires summing over all possible $w_{t'}$, which is computationally expensive. Typically the skipgram estimation problem is solved by negative sampling (Mikolov, Sutskever, et al., 2013), but this does not yield properly normalized probabilities. We therefore turn instead to noise contrastive estimation (NCE; Gutmann & Hyvärinen, 2010), which makes it possible to estimate the probability in Equation (1) without computing the normalization term (Mnih & Kavukcuoglu, 2013).

Suppose that the observed data are 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

$$J = \sum_t \log \Pr(D=1|w_t, w_{t'}) + \sum_{j=1}^{k} \log \Pr\left(D=0|w_t, \tilde{w}^{(j)}\right), \tag{2}$$

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

Now let us define the probability,

$$\Pr(D=1|, w_t w_{t'}) = \frac{P(w_{t'}|D=1, w_t)\Pr(D=1)}{P(w_{t'}|D=1, w_t)\Pr(D=1) + P(w_{t'}|D=0)\Pr(D=0)}, \tag{3}$$

$$= \frac{P(w_{t'}|w_t)}{P(w_{t'}|w_t) + kP_n(w_{t'})}, \tag{4}$$

$$= \left(1 + k\frac{P_n(w_{t'})}{P(w_{t'}|w_t)}\right)^{-1}, \tag{5}$$

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

$$\approx \sigma\left(\mathbf{v}_{w_{t'}} \cdot \mathbf{u}_{w_t} - \log\left(kP_n(w_{t'})\right)\right), \tag{7}$$

where $\sigma$ 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 (7) because the NCE objective is approximately "self-normalizing" when $P_n$ has positive support over all $w \in$ (Mnih & Kavukcuoglu, 2013). We then maximize Equation (2) by gradient ascent, which yields embeddings that are asymptotically equivalent to the optimizers of Equation (1) (Gutmann & Hyvärinen, 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.

The skipgram model is not identifiable: any permutation of the dimensions of the input and output

embeddings will yield the same result so many parameterizations of the model are observationally equivalent. For this reason, embeddings that are learned independently across multiple corpora must be aligned before their similarity can be quantified. To reconcile the input embeddings between the corpora $\mathcal{W}^{(\text{old})}$ and $\mathcal{W}^{(\text{new})}$ and make them comparable across the two corpora, we follow Hamilton et al. (2016b) and apply the Procrustes method (Gower & Dijksterhuis, 2004) to identify an orthogonal projection $\mathbf{Q}$ that minimizes the Frobenius norm $\|\mathbf{QU}^{(\text{old})} - \mathbf{U}^{(\text{new})}\|_F$, where $\|\mathbf{X}\|_F = \sqrt{\sum_{i,j} x_{i,j}^2}$.

### 2.1.1 | 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 (Levy & Goldberg, 2014; Sagi, Kaufmann, & Clark, 2011). As a result, the list of near neighbors can change across multiple runs (Hellrich & Hahn, 2016). Nonetheless, we chose NCE because the resulting embeddings outperformed alternatives on intrinsic word similarity benchmarks (Luong, Socher, & Manning, 2013). Our robustness checks indicated that the method identified similar sets of semantic innovations across multiple runs.

## 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. We therefore follow the local second-order approach proposed by Hamilton, Leskovec, and Jurafsky (2016a). First, for each word, we form the union of the sets of a word's near-neighbors ($n = 50$) in the "old" and "new" periods. Next, we compute the similarity of the word's embedding to the embeddings for members of this set, 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.

## 2.3 | Situating usages with respect to semantic change

Given a set of semantic innovations $\subset$, our main methodological innovation is to situate usage with respect to semantic changes. Each usage of an innovation $w^* \in$ 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, we compute the sum,

$$r_{w^*,i} = \sum_{t:w_{i,t}=w^*} \sum_{\substack{j \geq -k \\ j \leq k \\ j \neq 0}} \log \frac{P^{(\text{new})}\left(w_{i,t+j} \middle| w^*\right)}{P^{(\text{old})}\left(w_{i,t+j} \middle| w^*\right)}. \tag{8}$$

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. Note that the document $i$ need not necessarily be in either $\mathcal{W}^{(\text{old})}$ or $\mathcal{W}^{(\text{new})}$. Substituting the form of probability from Equation (1) and simplifying further, the log-likelihood ratio reduces to

$$\begin{aligned} r_{w^*,i} = \sum_{t:w_{i,t}=w^*} \sum_{\substack{j \geq -k \\ j \leq k \\ j \neq 0}} & \mathbf{v}_{w_{i,t+j}}^{(\text{new})} \cdot \mathbf{u}_{w^*}^{(\text{new})} - Z_{w^*}^{(\text{new})} \\ & - \mathbf{v}_{w_{i,t+j}}^{(\text{old})} \cdot \mathbf{u}_{w^*}^{(\text{old})} + Z_{w^*}^{(\text{old})}, \end{aligned} \tag{9}$$

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

## 2.4 | Aggregating to document scores

Given a set of innovations $\mathcal{S} \subset \mathcal{V}$, for each document $i$ we obtain 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 $\mu$ is the mean score for documents published in the same year, and $\sigma$ 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, we ensure that the progressiveness metric does not inherently favor older or newer texts.

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 number 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 Appendices A and B for this article.

## 3 | DATA

We empirically validate our approach on two document collections: documents representing *legal opinions* in federal courts of the United States of America (Lerman, Hodas, & Wu, 2017),[2] and the DBLP collection of *computer science abstracts* (Ley, 2002).[3] These data sets were chosen because they include timestamps as well as citation information, making it possible to link semantic innovation with influence in a citation network.

### 3.1 | 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. We obtained all legal opinions by using the bulk API of a publicly available service.[4] These opinions span over 400 courts, multiple centuries, and have a broad jurisdictional coverage.

### 3.2 | Scientific abstracts

The abstracts from DBLP were obtained from ArnetMiner,[5] a service that has released multiple versions of this data with the latest papers since 2010 (Sinha et al., 2015; Tang et al., 2008). We chose the latest version (v10) from their collection.

### 3.3 | Metadata

Both data sets 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 1.

**TABLE 1** Descriptive summary of the two datasets

| 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 |

## 4 | IDENTIFYING SEMANTIC INNOVATIONS

We now describe the steps taken to create a list of semantic innovations in these data sets. These innovations are then used to score every document for its progressiveness.

### 4.1 | 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 & Baldwin, 2012). In both collections, we converted the text to lowercase before proceeding, and employed *spaCy* for tokenization.[6]

### 4.2 | 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.[7] We ignored tokens with frequency below a predetermined threshold: 5 for the abstracts and 10 for the larger data set of legal 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).

### 4.3 | Postprocessing

After estimating the embeddings, semantic innovations were identified using the technique described in Discovering Semantic Innovations. The number of nearest neighbors used for the computation of the metric was set to 50.

#### 4.3.1 | 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 then 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 pretrained tagger provided by *spaCy*.[8] If a term is either (a) labeled 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.

#### 4.3.2 | Abbreviations

In the data set of scientific abstracts, the mention of names is rare, but abbreviations pose a similar challenge. We identify abbreviations using a similar heuristic procedure as described earlier: 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 earlier, 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 data set, we retain a list of the 1000 terms that underwent the most substantial semantic changes, as measured by overlap in their semantic neighborhood (described earlier). 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 described later.

## 5 | INNOVATIONS AND INNOVATORS

### 5.1 | Semantic changes

A few prominent semantic innovations are listed in Table 2. The innovations in the legal opinions corpus we discover span multiple domains, including financial (e.g., *laundering*, which earlier exclusively meant washing), sociopolitical (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

**TABLE 2** Examples of semantic innovations identified by our method for both the data sets

| Doc. Type | Innovations | Old usage | New usage | Example document with new usage |
|---|---|---|---|---|
| Legal | Laundering | *Laundering* clothes | *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 and Tractinsky (2006) |
| | Surf | *Surf* the Internet | Descriptor *surf* | Bay, Ess, Tuytelaars, and Van Gool (2008) |
| | Android | Intelligent *android* | Google's *android* | Shabtai et al. (2010) |

the Internet). Our analysis also independently discovers semantic changes in words like *cellular* and *asylum*, which have previously been identified as semantic changes in other corpora (Hamilton et al., 2016a, 2016b; Kulkarni et al., 2015).

In the scientific domain, a common source of semantic innovation is through the use of abbreviations (recall that the filtering steps in the previous section exclude only rare 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 with 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.

## 5.2 | Leading documents

Two examples of legal opinions at the leading edge of change according to our metrics 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:

- ...2-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.

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 2 shows the evolution of the semantic innovations, which approximately aligns with the leading documents that our method discovered.

## 5.3 | Computational requirements

Our method to score the progressiveness of documents can effectively leverage computational resources without putting a heavy burden on them. Following Equation (9), we calculate the normalization terms for a specific change just once for the entire corpus to speed up the calculation considerably. Moreover, the progressiveness calculation is also embarrassingly parallelizable; that is, scores for different documents can be computed in parallel. We calculated the progressiveness scores of millions of documents on an 80-processor machine with 256 GB memory in just over a couple of hours.

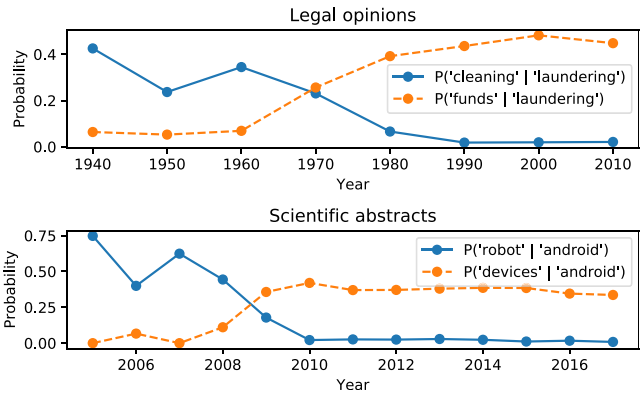


**FIGURE 2** Examples of semantic changes identified by the method. In the upper time series, the meaning of the term *laundering* evolves to include money laundering, shown here as the increase in the conditional probability of seeing the term *laundering* given that *funds* appears in a document, in 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) [Color figure can be viewed at wileyonlinelibrary.com]

## 6 | INNOVATION AND INFLUENCE

While the examples mentioned in the earlier section are suggestive of the validity of our method for identifying innovations and innovators, additional validation is necessary. Lacking large-scale manual annotations for the semantic progressiveness of legal opinions or scientific abstracts, we instead measure *influence*, 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 we do not hypothesize a one-way causal relationship from semantic innovation to influence. 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—for example, 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, Ng, & Jordan, 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.

### 6.1 | Univariate analysis

Figure 3 shows the number of citations for each quartile of our progressiveness measure, indicating a steady increase in both data sets. This figure excludes documents that do not include any of the terms identified as having changing semantics. We also exclude 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.

### 6.2 | Multivariate analysis

There are many factors that drive citation counts, such as age, length, and content (Fowler, Johnson, Spriggs, Jeon, & Wahlbeck, 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, we formulate the problem 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. As the number of citations is a count variable, we fit a Poisson regression model.[9]

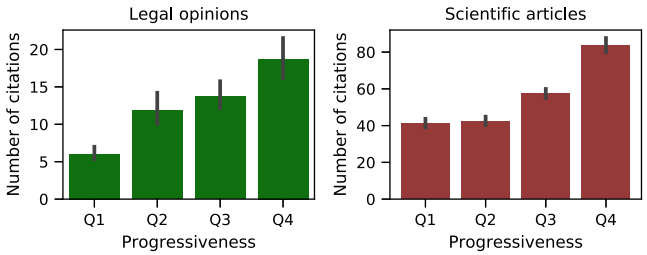


**FIGURE 3** The univariate relationship between the number of citations and our measure of semantic progressiveness. For both the legal opinions and the scientific articles, the citations increase for more progressive documents [Color figure can be viewed at wileyonlinelibrary.com]

### 6.3 | Regression models

To assess the relevance of semantic progressiveness, we compare against two baseline models, which include covariates that capture structural information about each document: the number of outgoing references that a document makes; its age; its length, operationalized as the number of unique types; and the number of authors for the document (available only for scientific articles). The baseline also incorporates a lightweight model of document content, to account for the fact that some topics may get cited more than others. Specifically, we fit a bag-of-words regression model on a small subset of documents (similar to Yogatama et al., 2011), and use its prediction as a covariate in the multivariate regression. We refer to this covariate as BoWs. This baseline is referred to as M1.

The second baseline, M2, includes all the covariates from M1, and an additional covariate for the number of unique semantic innovations present in the document. This is aimed to tease out the effect of the *presence* of words with changing semantics from the extent to which the document employs the more contemporary meaning, as captured by our measure of semantic progressiveness. We refer to this covariate as # Innovs.

To test the effect of semantic progressiveness, we create two experimental models, M3 and M4, which use the $z$-scores described earlier. In M3, the $z$-score is included

as a raw value; in M4 it is binned into quartiles. Note that for M4, the bottom quartile (Q1) receives a coefficient of zero by default, so that the model is not underdetermined.

We compare these models by goodness of fit, which is a standard technique from quantitative social science (Greene, 2003). We compute the log-likelihood for each model; under the null hypothesis that the more complex model is no better than the baseline, the log-likelihood ratio has a $\chi^2$ distribution with degrees of freedom equal to the number of parameters in the more expressive model. If the observed log-likelihood ratio is unlikely to arise under this distribution, then we can reject the null hypothesis. This approach is similar in spirit to the Akaike Information Criterion, which also penalizes the log-likelihood by the number of parameters. We can also measure the effect size by examining the regression coefficients: the value of each coefficient corresponds to an increase in the log of the expected number of citations under the Poisson model.

## 6.4 | Results

The regressions reveal a strong relationship between semantic progressiveness and citation count. For the scientific abstracts (Table 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). 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 in Table 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.

Since age is a strong predictor of citation count for both the data sets, we also compared the distribution of publication years in each quartile to rule out the possibility that the higher quartiles could be acting as another proxy for age. Using a Kolmogorov–Smirnov test, we tested whether the distributions in each pair of quartiles were different, with the null hypothesis being that they were statistically equivalent. We found that statistical equivalence could not be ruled out (on scientific abstract collection the $p$ values ranged from 0.3 to 0.92, and on the court's opinion collection the $p$ values ranged from 0.1 to 0.18), meaning that the semantic progressiveness scoring scheme does not discriminately favor old or recent documents. Overall, these results indicate that our 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.

**TABLE 3** Poisson regression analysis of citations to scientific abstracts

| Predictors | M1 | M2 | M3 | M4 |
|---|---|---|---|---|
| Constant | 1.983 | 1.943 | 2.032 | 1.770 |
| | (0.001) | (0.001) | (0.001) | (0.001) |
| Outdegree | 0.009 | 0.009 | 0.009 | 0.009 |
| | (0.000) | (0.000) | (0.000) | (0.000) |
| # Authors | 0.055 | 0.054 | 0.054 | 0.054 |
| | (0.000) | (0.000) | (0.000) | (0.000) |
| Age | 0.079 | 0.079 | 0.078 | 0.073 |
| | (0.000) | (0.000) | (0.000) | (0.000) |
| Length | 0.002 | 0.002 | 0.002 | 0.002 |
| | (0.000) | (0.000) | (0.000) | (0.000) |
| BoWs | 0.000 | 0.000 | 0.000 | 0.000 |
| | (0.000) | (0.000) | (0.000) | (0.000) |
| # Innovs | | 0.028 | −0.010 | −0.034 |
| | | (0.000) | (0.000) | (0.000) |
| Prog. | | | 0.137 | |
| | | | (0.000) | |
| Prog. Q2 | | | | 0.179 |
| | | | | (0.001) |
| Prog. Q3 | | | | 0.431 |
| | | | | (0.001) |
| Prog. Q4 | | | | 0.698 |
| | | | | (0.001) |
| Log Lik. | −13.07 | −13.06 | −12.93 | **−12.82** |

*Note:* 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. The significance of bold value is threshold was 0.001.

# 7 | RELATED WORK

## 7.1 | Language change

Language change has been a topic of great general interest. Much of the early computational work on language

**TABLE 4** Poisson regression analysis of citations to legal documents

| Predictors | M1 | M2 | M3 | M4 |
|---|---|---|---|---|
| Constant | 1.614 | 1.421 | 1.476 | 1.168 |
| | (0.003) | (0.004) | (0.004) | (0.006) |
| Outdegree | 0.022 | 0.020 | 0.021 | 0.020 |
| | (0.000) | (0.000) | (0.000) | (0.000) |
| Age | 0.009 | 0.011 | 0.010 | 0.010 |
| | (0.000) | (0.000) | (0.000) | (0.000) |
| Length | 0.000 | −0.000 | −0.000 | −0.000 |
| | (0.000) | (0.000) | (0.000) | (0.000) |
| BoWs | −0.000 | −0.000 | −0.000 | −0.000 |
| | (0.000) | (0.000) | (0.000) | (0.000) |
| # Innovs | | 0.054 | 0.045 | 0.042 |
| | | (0.001) | (0.001) | (0.001) |
| Prog. | | | 0.094 | |
| | | | (0.001) | |
| Prog. Q2 | | | | 0.384 |
| | | | | (0.007) |
| Prog. Q3 | | | | 0.382 |
| | | | | (0.007) |
| Prog. Q4 | | | | 0.470 |
| | | | | (0.007) |
| Log Lik. | −415,195 | −410,601 | **−406,843** | −408,031 |

*Note:* Each column indicates a model, each row indicates a predictor, and each cell contains the coefficient and, in parentheses, its standard error. The significance of bold value is threshold was 0.001.

change focused on tracking the *frequency* of lexical items, rather than their meaning. Michel et al. (2011) track changes in word frequency in large books corpora, and link these changes to social-cultural trends and events. Danescu-Niculescu-Mizil, West, Jurafsky, Leskovec, and Potts (2013) compare the adoption rates of words between community members. Eisenstein, O'Connor, Smith, and Xing (2014) track the diffusion of new words over geographical regions, and Goel et al. (2016) model diffusion across social networks. Measures of linguistic progressiveness in this line of work are also based on frequency and other dependent statistics, such as cross entropy (Danescu-Niculescu-Mizil et al., 2013) or *tf-idf* (Kelly, Papanikolaou, Seru, & Taddy, 2018). More recently, several methods have been proposed to learn diachronic word embeddings as a means to track language change at a finer semantic level. These methods include matrix decomposition (e.g., Yao, Sun, Ding, Rao, & Xiong, 2018), Bayesian inference (e.g., Bamler & Mandt, 2017; Frermann & Lapata, 2016; Wijaya & Yeniterzi, 2011), and neural word embeddings (e.g., Hamilton et al., 2016b; Kim, Chiu, Hanaki, Hegde, & Petrov, 2014; Kulkarni et al., 2015; Rosenfeld & Erk, 2018). Diachronic word embeddings have shown success in identifying linguistic (Hamilton et al., 2016a) and sociocultural changes over time (Garg, Schiebinger, Jurafsky, & Zou, 2018). Two surveys review the existing research on diachronic language change through word embeddings (Kutuzov et al., 2018; Tahmasebi, Borin, & Jatowt, 2018). Contextual word representations (e.g., Devlin, Chang, Lee, & Toutanova, 2019), which have produced state-of-the-art results in natural language processing, have also been utilized to identify semantic changes (Giulianelli, Del Tredici, & Fernández, 2020). However, despite these successes, prior work has not provided methods to identify the documents at the forefront of semantic change. Our work specifically addresses this gap.

Another body of work has used topic modeling to study changes over time (e.g., Blei & Lafferty, 2006; Mimno, 2012; Wang & McCallum, 2006). Of particular relevance is the use of topical changes in scientific literature to discover documents with the most scholarly impact (Gerrish & Blei, 2010; Hall, Jurafsky, & Manning, 2008). We argue that these approaches are complementary. While topic models provide a macrolevel view of the concerns and interests of a set of writers, word embeddings provide a more fine-grained perspective by demonstrating shifts in meaning of individual terms. Topic models are centered at the document level, and so make it easy to identify innovators; our work extends this capability to embedding-based analysis of semantic change.

## 7.2 | Citation impact

The number of citations a document receives has long been used as a proxy for the impact and influence of scientific articles (Fortunato et al., 2018) and legal opinions (Fowler et al., 2007), as well as researchers and scientific trends (Börner, Maru, & Goldstone, 2004). Dynamic models capturing the mechanics of attention have been modestly successful in predicting long-term scientific impact (Wang, Song, & Barabási, 2013). Other models accounting for changing language have been used to identify important new topics (Börner et al., 2004) or to estimate the influence of papers on one another (Dietz, Bickel, & Scheffer, 2007). In a different domain, progressiveness as measured in terms of textual dissimilarity with past patents and textual similarity with future patents is shown to be predictive of future citations of a

patent (Kelly et al., 2018). Our quantitative insights in this work are similar to that of Kelly et al. (2018) but our measure of language change is more fine-grained and is based on semantic changes, instead of textual similarity that can depend on topics or events.

## 8 | CONCLUSION

This article 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). Our current method of identifying semantically progressive usage is limited to words. For future work, we are interested in extending the current method beyond words to phrases and grouping phrases and words with semantic similarity together—for example, this would enable the word *LDA* and the phrase *Latent Dirichlet Allocation* to have similar embeddings and potentially improve the results for progressive usage identification. In future work, we are also interested to assess how semantically progressive documents are received by their audiences and to explore semantic change as a site of linguistic contestation. For example, recent work has linked diachronic word embeddings to gender and ethnic stereotypes in large-scale data sets of books (Garg et al., 2018). Our method could link author and audience covariates with the documents that led and trailed changes in these stereotypical associations, providing new insight on these historical trends. Finally, contemporaneous work has demonstrated the use of contextualized embeddings, highly powerful word representations, in detecting semantic changes (Giulianelli et al., 2020). In future, we are interested to extend contextual embeddings to measure semantic progressiveness.

## ENDNOTES

[1] We have released the code and the word embeddings for both the data sets at https://github.com/sandeepsoni/semantic-progressiveness/

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

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

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

[5] https://aminer.org

[6] https://spacy.io/

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

[8] 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.

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

## REFERENCES


Antoniak, M., & Mimno, D. (2018). Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, *6*, 107–119.

Bamler, R., & Mandt, S. (2017). Dynamic word embeddings. *Proceedings of the 34th International Conference on Machine Learning*, *70*, 380–389.

Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, *110*(3), 346–359.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*, 113–120. https://doi.org/10.1145/1143844.1143859

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Börner, K., Maru, J. T., & Goldstone, R. L. (2004). The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5266–5273.

Boyack, K. W., & Klavans, R. (2014). Creation of a highly detailed, dynamic, global model and map of science. *Journal of the Association for Information Science and Technology*, *65*(4), 670–685.

Bruni, E., Boleda, G., Baroni, M., & Tran, N.-K. (2012). Distributional semantics in technicolor. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long papers*, *1*, 136–145.

Burdick, L., Kummerfeld, J. K., & Mihalcea, R. (2018). Factors influencing the surprising instability of word embeddings. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*, 2092–2102.

Cameron, A. C., & Trivedi, P. K. (1990). Regression-based tests for overdispersion in the poisson model. *Journal of Econometrics*, *46*(3), 347–364.

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., & Potts, C. (2013). No country for old members: User lifecycle and linguistic change in online communities. *Proceedings of the 22nd International Conference on World Wide Web*, 307–318. https://doi.org/10.1145/2488388.2488416

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). 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*) (pp. 4171–4186).Association for Computational LinguisticsMinneapolis, Minnesota

Dietz, L., Bickel, S., & Scheffer, T. (2007). Unsupervised prediction of citation influences. In *Proceedings of the 24th International Conference on Machine Learning* (pp. 233–240).

Ding, Y. (2011). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of Informetrics*, *5*(1), 187–203.

Dyer, C. (2014). Notes on noise contrastive estimation and negative sampling. *arXiv* arXiv:1410.8251.

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2014). Diffusion of lexical change in social media. *PLoS One*, *9*(11), 1–13.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojevic, S., ... Barabási, A.-L. (2018). Science of science. *Science*, *359*(6379), eaao0185.

Fowler, J. H., Johnson, T. R., Spriggs, J. F., Jeon, S., & Wahlbeck, P. J. (2007). Network analysis and the law: Measuring the legal importance of precedents at the US Supreme court. *Political Analysis*, *15*(3), 324–346.

Frermann, L., & Lapata, M. (2016). A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, *4*, 31–45.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), 3635–3644.

Gerow, A., Hu, Y., Boyd-Graber, J., Blei, D. M., & Evans, J. A. (2018). Measuring discursive influence across scholarship. *Proceedings of the National Academy of Sciences*, *115*(13), 3308–3313.

Gerrish, S. M., & Blei, D. M. (2010). A language-based approach to measuring scholarly impact. In *Proceedings of the 27th International Conference on Machine Learning*. (pp. 375–382).

Giulianelli, M., Del Tredici, M., & Fernández, R. (2020). Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 3960–3973

Goel, R., Soni, S., Goyal, N., Paparrizos, J., Wallach, H., Diaz, F., & Eisenstein, J. (2016). The social dynamics of language change in online networks. In *International conference on social informatics*. (pp. 41–57).

Gower, J. C., & Dijksterhuis, G. B. (2004). *Procrustes problems* (Vol. 30). Oxford, UK: Oxford University Press.

Greene, W. H. (2003). *Econometric analysis* (5th ed.). Upper Saddle River, NJ: Pearson Education.

Gutmann, M., & Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics* (pp. 297–304).

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the history of ideas using topic models. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. (pp. 363–371).

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016a). Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. (pp. 2116–2121).

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016b). 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)*. (pp. 1489–1501).

Hassenzahl, M., & Tractinsky, N. (2006). User experience - a research agenda. *Behaviour & Information Technology*, *25*(2), 91–97.

Hellrich, J., & Hahn, U. (2016). Bad company—neighborhoods in neural embedding spaces considered harmful. In *Proceedings of the 26th International Conference on Computational Linguistics*. (pp. 2785–2796).

Kelly, B. T., Papanikolaou, D., Seru, A., & Taddy, M. (2018). Measuring technological innovation over the long run. *NBER Working Paper* (w25266).

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., & Petrov, S. (2014). Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. (pp. 61–65).

Kulkarni, V., Al-Rfou, R., Perozzi, B., & Skiena, S. (2015). Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*. (pp. 625–635).

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*. (pp. 1384–1397).

Lerman, K., Hodas, N. O., & Wu, H. (2017). Bounded rationality in scholarly knowledge discovery. *arXiv*. arXiv:1710.00269 Retrieved fromhttp://arxiv.org/abs/1710.00269

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. *Proceedings of the 27th International Conference on Neural Information Processing Systems*, *2*, 2177–2185.

Ley, M. (2002). The DBLP computer science bibliography: Evolution, research issues, perspectives. In *International Symposium on String Processing and Information Retrieval* (pp. 1–10).

Lui, M., & Baldwin, T. (2012). langid.py: An off-the-shelf language identification tool. . In *Proceedings of the ACL 2012 System Demonstrations* (pp. 25–30).

Luong, T., Socher, R., & Manning, C. (2013). Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*. (pp. 104–113).

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P.,... others (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331* (6014), 176–182.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*. (pp. 3111–3119).

Mimno, D. (2012). Computational historiography: Data mining in a century of classics journals. *Journal on Computing and Cultural Heritage*, *5*(1), 3.

Mnih, A., & Kavukcuoglu, K. (2013). Learning word embeddings efficiently with noise-contrastive estimation. . In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2* (pp. 2265–2273).

Nallapati, R. M., Ahmed, A., Xing, E. P., & Cohen, W. W. (2008). Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. (pp. 542–550).

Rosenfeld, A., & Erk, K. (2018). 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)* (pp. 474–484).

Sagi, E., Kaufmann, S., & Clark, B. (2011). Tracing semantic change with latent semantic analysis. *Current Methods in Historical Semantics*, *73*, 161–183.

Shabtai, A., Fledel, Y., Kanonov, U., Elovici, Y., Dolev, S., & Glezer, C. (2010). Google android: A comprehensive security assessment. *IEEE Security & Privacy*, *8*, 35–44.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J. P., & Wang, K. (2015). An overview of Microsoft academic service (MAS) and applications. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 243–246).

Tahmasebi, N., Borin, L., & Jatowt, A. (2018). Survey of computational approaches to diachronic conceptual change. *arXiv* arXiv:1811.06278.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). Arnetminer: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 990–998).

Traugott, E. C., & Dasher, R. B. (2001). *Regularity in semantic change* (Vol. 97), Cambridge, England: Cambridge University Press.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Van Opijneni, M. (2012). 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*, p. 95).

Van Raan, A. (1997). Scientometrics: State-of-the-art. *Scientometrics*, *38*(1), 205–218.

Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying long-term scientific impact. *Science*, *342*(6154), 127–132.

Wang, X., & McCallum, A. (2006). 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*. (pp. 424–433).

Weinreich, U., Labov, W., & Herzog, M. (1968). Empirical foundations for a theory of language change. *Directions for Historical Linguistics*, *58*, 97–188.

Wijaya, D. T., & Yeniterzi, R. (2011). Understanding semantic change of words over centuries. In *Proceedings of the 2011 International Workshop on Detecting and Exploiting Cultural Diversity on the Social Web*. (pp. 35–40).

Xu, J., Bu, Y., Ding, Y., Yang, S., Zhang, H., Yu, C., & Sun, L. (2018). Understanding the formation of interdisciplinary research from the perspective of keyword evolution: A case study on joint attention. *Scientometrics*, *117*(2), 973–995.

Yan, E. (2015). Research dynamics, impact, and dissemination: A topic-level analysis. *Journal of the Association for Information Science and Technology*, *66*(11), 2357–2372.

Yao, Z., Sun, Y., Ding, W., Rao, N., & Xiong, H. (2018). Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (pp. 673–681).

Yogatama, D., Heilman, M., O'Connor, B., Dyer, C., Routledge, B. R., & Smith, N. A. (2011). Predicting a scientific community's response to an article. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 594–604).

**How to cite this article:** Soni S, Lerman K, Eisenstein J. Follow the leader: Documents on the leading edge of semantic change get more citations. *J Assoc Inf Sci Technol*. 2021;72:478–492. https://doi.org/10.1002/asi.24421

## APPENDIX A

### 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 & Mimno, 2018; Burdick, Kummerfeld, & Mihalcea, 2018). We ran our pipeline of learning word embeddings, identifying semantic innovations, and measuring the semantic progressiveness of documents for different random initialization.

### 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 initializations on benchmark test sets to verify that our method is quite stable. Specifically, we evaluate the

**TABLE A1** Accuracy (in %) of word embeddings on the analogy test set (Mikolov, Chen, Corrado, & Dean, 2013

| Runs | CourtListener Early | CourtListener Later | DBLP Early | DBLP 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 A2** Spearman correlation of word embeddings on the word similarity test set (Bruni, Boleda, Baroni, & Tran, 2012)

| Runs | CourtListener Early | CourtListener Later | DBLP Early | DBLP 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 |

quality of the word embeddings on analogy and word similarity tasks for three runs, each differing in the initialization point. The results are in Tables A1 and A2, respectively.

### 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 in Table A3 and for the DBLP collection in Table A4.

**TABLE A3** Top semantic changes across different runs for Courtlistener text collection

| 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* |

*Note:* Six changes appear in the top 10 across all three runs, as shown in italics.

**TABLE A4** Top semantic changes across different runs for DBLP text collection

| 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* |

*Note:* Nine changes appear in the top 10 across all three runs, as shown in italics.

**TABLE A5** 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 |

### 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 A5 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.

## APPENDIX B

### Alternative measurement of semantic progressiveness

The results from the multivariate regressions using an alternative measure of semantic progressiveness.

**TABLE B1** Poisson regression analysis of citations to scientific abstracts

| | M1 | M2 | M3 |
|---|---|---|---|
| Constant | 2.078 | 2.011 | 2.008 |
| | (0.001) | (0.001) | (0.001) |
| Outdegree | 0.010 | 0.010 | 0.010 |
| | (0.000) | (0.000) | (0.000) |
| # Authors | 0.024 | 0.024 | 0.024 |
| | (0.000) | (0.000) | (0.000) |
| Age | 0.074 | 0.077 | 0.076 |
| | (0.000) | (0.000) | (0.000) |
| Length | 0.002 | 0.002 | 0.002 |
| | (0.000) | (0.000) | (0.000) |
| BoWs | 0.000 | 0.000 | 0.000 |
| | (0.000) | (0.000) | (0.000) |
| Prog. | | 0.049 | |
| | | (0.000) | |
| Prog. Q2 | | | 0.105 |
| | | | (0.001) |
| Prog. Q3 | | | 0.045 |
| | | | (0.001) |
| Prog. Q4 | | | 0.137 |
| | | | (0.001) |
| Log Lik. | −12.923 | **−12.891** | −12.912 |

*Note:* 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.

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 B1 contains the results for the DBLP collection and the Table B2 contains the results for the collection of court opinions.

**TABLE B2** Poisson regression analysis of citations to legal documents

| | M1 | M2 | M3 |
|---|---|---|---|
| Constant | 1.612 | 1.515 | 0.963 |
| | (0.003) | (0.004) | (0.007) |
| Outdegree | 0.019 | 0.020 | 0.019 |
| | (0.000) | (0.000) | (0.000) |
| Age | 0.011 | 0.012 | 0.017 |
| | (0.000) | (0.000) | (0.000) |
| Length | 0.000 | 0.000 | 0.000 |
| | (0.000) | (0.000) | (0.000) |
| BoWs | 0.000 | 0.000 | 0.000 |
| | (0.000) | (0.000) | (0.000) |
| Prog. | | 0.051 | |
| | | (0.001) | |
| Prog. Q2 | | | 0.577 |
| | | | (0.006) |
| Prog. Q3 | | | 0.615 |
| | | | (0.007) |
| Prog. Q4 | | | 0.745 |
| | | | (0.008) |
| Log Lik. | −429,096 | −427,724 | **−423,474** |

*Note:* Each column indicates a model, each row indicates a predictor, and each cell contains the coefficient and, in parentheses, its standard error. The significance of bold value is threshold is 0.001.