Gallagher et al. *EPJ Data Science* (2021) 10:4
https://doi.org/10.1140/epjds/s13688-021-00260-3

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

# Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts

Ryan J. Gallagher[1*], Morgan R. Frank[2,3,4], Lewis Mitchell[5], Aaron J. Schwartz[6,7,8], Andrew J. Reagan[9], Christopher M. Danforth[6,10] and Peter Sheridan Dodds[6,10]

*Correspondence:
gallagher.r@northeastern.edu
[1]Network Science Institute, Northeastern University, 02115 Boston, MA, USA
Full list of author information is available at the end of the article

**Abstract**

A common task in computational text analyses is to quantify how two corpora differ according to a measurement like word frequency, sentiment, or information content. However, collapsing the texts' rich stories into a single number is often conceptually perilous, and it is difficult to confidently interpret interesting or unexpected textual patterns without looming concerns about data artifacts or measurement validity. To better capture fine-grained differences between texts, we introduce generalized word shift graphs, visualizations which yield a meaningful and interpretable summary of how individual words contribute to the variation between two texts for any measure that can be formulated as a weighted average. We show that this framework naturally encompasses many of the most commonly used approaches for comparing texts, including relative frequencies, dictionary scores, and entropy-based measures like the Kullback–Leibler and Jensen–Shannon divergences. Through a diverse set of case studies ranging from presidential speeches to tweets posted in urban green spaces, we demonstrate how generalized word shift graphs can be flexibly applied across domains for diagnostic investigation, hypothesis generation, and substantive interpretation. By providing a detailed lens into textual shifts between corpora, generalized word shift graphs help computational social scientists, digital humanists, and other text analysis practitioners fashion more robust scientific narratives.

**Keywords:** Text as data; Data visualization; Word shift graphs; Sentiment analysis; Computational social science; Digital humanities; Natural language processing; Information theory

## 1 Introduction

News articles, audio transcripts, medical records, digitized archives, virtual libraries, computer logs, online memes, open-ended questionnaires, legislative proceedings, political manifestos, fan fiction, and poetry collections are just some of the many large-scale data sources that are readily available as text data [1–3]. Computational methods help funnel what would be an otherwise overwhelming fire hose of raw text into coherent streams of social information [3]. Social media text has allowed us to ask about the emotional pulse of large populations [4, 5], the subtle adoption of community language by new members

[6], and the role of social bots in instigating political dialogue [7]. Digitized archives and collections have made it possible to observe the deliberative evolution of the French revolution [8], the lifespans of words across centuries [9, 10], and the social roles of characters in works of fiction [11]. Song lyrics let us infer the latent emotions associated with musical chords [12], legislative corpora show us the reuse of statutes across jurisdictions [13], and transcribed police body camera footage echos lived experiences of racial disparities in officer respect [14]. Text as data fundamentally expands the number of social questions that we can ask across many different domains.

Computational methods for dealing with texts are abundant, but at the backbone of many of them is an intuitive concept: the weighted average. Weighted averages are a convenient tool because they are mathematically simple—it is easy to draw pairwise comparisons between texts by averaging over them in their entirety [15–17] or measure temporal trajectories by repeatedly averaging over time [8–10, 18, 19]. For example, sentiment analysis, one of the most popular applications of weighted averages, can be used to compare the "happiness" expressed online by different parts of the United States [5, 20]. First, each word is assigned a score based on how much happiness is associated with it. Then, for each different geographic region, the average happiness is computed by summing how often different words appear and weighting them by their happiness scores. Similarly, the expressed happiness of an online population can observed over time by repeatedly taking the weighted average over all the text from each successive day, week or month [4]. Beyond sentiment, domain knowledge [21, 22] and other social scientific constructs like morality [23], respect [14], and hatefulness [24, 25] can also be integrated through weights. This makes it easy to adapt average-based methods to new situations and focus them on particular questions of interest.

However, the simplicity of the weighted average is often one of its most significant drawbacks. Collapsing texts down to a single number introduces serious concerns about measurement validity because it is not always clear a mere weighted average can capture complex social phenomena [3, 21, 26, 27]. Even if one accepts a particular weighted average as a conceptually valid measurement, that measure can still vary in unanticipated ways. The sheer abundance of language underlying computational text models [28, 29] can cause a given measure to rise or fall due to the frequent appearance of a single set of key words [30] or an unexpected combination of frequent and less frequent words [16, 31]. Further, the relative weights of those words may be highly context dependent [3, 26]—regional dialects [32, 33], variations in slang [34], and other domain-specific usage [26, 35, 36] all affect how appropriate it is to compare across weighted averages. Even if contextual weights can be derived for different sets of text, there are limited tools for comparing weighted averages beyond their aggregate value. While theory can provide guidance at times, it is a perilous path towards reliably interpreting text data if we do not have methods for interpreting the averages themselves.

We contend that these concerns can and should be addressed by systematically quantifying *which* words contribute to the differences between two texts, and, importantly, *how* they do so. To this end, we propose *generalized word shift graphs*, horizontal bar charts which provide word-level explanations of how and why two texts differ across any measure derived from a weighted average. The framework that we propose generalizes previous formulations of word shifts [4, 37] to account for how a word changes in both relative frequency and measurement, allowing us to unify a wide range of common measures under the same methodological banner, including dictionary scores, Shannon entropy, the Kullback–Leibler divergence, the Jensen–Shannon divergence, generalized entropies, and any other measure that can be written as a weighted average or difference in weighted averages.

Through a number of case studies, we show that generalized word shift graphs address many of the aforementioned issues: they unmask the internal workings of aggregate averages, enumerate exactly which words contribute to variation in a measure, account for context-dependent measurements across different settings, diagnose measurement issues during the research process, and provide an interpretable tool for validating, constructing, and presenting scientifically sound stories. These case studies span presidential speeches, classic novels, tweets from U.S. urban parks, social media platform changes, and employment diversity of labor markets, demonstrating the versatility of shift graphs across domains. We advocate for the use of generalized word shift graphs among computational social scientists, digital humanists, and other text analysis practitioners, and release open source code to encourage their uptake in the methodological toolkit for working with text as data. We make our open source code for constructing generalized word shift graphs available at https://github.com/ryanjgallagher/shifterator.

## 2 Pairwise comparisons between texts

We first present a number of measures that are representative of the many different ways that two different texts can be quantitatively juxtaposed. As we show more explicitly later, all of these can be written as a weighted average or difference in weighted averages. For each measure, we provide guidance on the questions that it is most capable of answering, and the benefits and limitations of applying the measure to draw out differences between texts.

Throughout the paper, we denote our two text corpora by $T^{(1)}$ and $T^{(2)}$. We consider the full vocabulary $\mathcal{T}$, composed of all the word types in either $\mathcal{T}^{(1)}$ or $\mathcal{T}^{(2)}$. Each word type $\tau$ in the vocabulary $\mathcal{T}$ appears with some frequency $f_\tau^{(i)}$ in each of the texts, where either $f_\tau^{(1)}$ or $f_\tau^{(2)}$ may be zero. We notate each type's normalized, relative frequency as $p_\tau^{(i)} = f_\tau^{(i)} / \sum_{\tau' \in \mathcal{T}} f_{\tau'}^{(i)}$. Unless otherwise specified, we use "word" to mean "word type," where a "word" may be any $n$-gram or phrase as defined by the vocabulary, and not necessarily just a unigram.

### 2.1 Relative frequency

One of the simplest and most common ways of identifying the most characteristic words of two texts is to compare how often each word appears in one text versus the other. That is, we can compute the difference in their relative frequencies,

$$p_\tau^{(2)} - p_\tau^{(1)}. \tag{1}$$

As we can see, if the difference is positive then the word is relatively more common in $T^{(2)}$, if it is negative then it is more common in $T^{(1)}$, and if it is zero then it is equally common in both texts. We can rank words by the magnitude of this difference to produce a list of words that distinguish the texts from one another.

Comparing the relative frequency of words is adequate for a cursory pass of two texts, but it is less attuned to identifying subtle, but characteristic differences between them.

Consider a word used frequently in both $T^{(1)}$ and $T^{(2)}$. Then the absolute difference $|p_\tau^{(2)} - p_\tau^{(1)}|$ has more potential for being large because $p_\tau^{(1)}$ and $p_\tau^{(2)}$ are themselves large. Yet, exactly because the word is frequently used, it is unlikely that the difference in usage will be surprising or substantively interesting. On the other hand, a less frequently used but more distinct word, can only have a difference as large as the maximum of $p_\tau^{(1)}$ and $p_\tau^{(2)}$, hindering its ability to rank highly. Comparing the relative frequencies of words puts more emphasis on differences between the most frequently used words, and less on the long, rich tail of word usage [28, 29] that may leave more lexical clues to what characterizes the texts.

## 2.2 Entropy

Shannon entropy accounts for both a word's relative frequency and its unexpectedness. If we let $P$ denote the entire normalized distribution of words in a text with vocabulary $\mathcal{T}$, then the (Shannon) entropy [38] is given by

$$H(P) = \sum_{\tau \in \mathcal{T}} p_\tau \log_2 \frac{1}{p_\tau}. \tag{2}$$

The entropy measures the unpredictability of a text: it is maximized if every word is equally likely to occur (i.e., $p_\tau = 1/N$ for all $N$ words in the vocabulary), and minimized if only one word is used (i.e., $p_\tau = 1$ for a single word $\tau$ and 0 for all others). At the word level, the factor $\log_2 1/p_\tau$ distinguishes a word's contribution to $H(P)$ from just its relative frequency $p_\tau$. This factor is known as a word's *surprisal*—a word is more surprising if it is used relatively less. Another way of interpreting the entropy then is as the average surprisal of a text.

To compare two texts, we can consider the difference in their entropies,

$$H\bigl(P^{(2)}\bigr) - H\bigl(P^{(1)}\bigr). \tag{3}$$

By considering the components of the sums, we can decompose the difference into the contribution from each word $\tau$,

$$\delta H_\tau = p_\tau^{(2)} \log_2 \frac{1}{p_\tau^{(2)}} - p_\tau^{(1)} \log_2 \frac{1}{p_\tau^{(1)}}. \tag{4}$$

Like relative frequencies, we can order words by their absolute contribution to obtain a ranked list of the words that are most characteristic of each text. Unlike relative frequencies, each word's surprisal weights it inversely to its frequency. Generalized, or Tsallis, entropies [39] introduce a tunable parameter to further control how much consideration is given to rare and common words [40–42] (see Materials and methods for details), and the Shannon entropy is a special limiting case that statistically balances between those that frequently and infrequently occur [40, 41]. Entropy has been particularly effective as an operationalization of diversity [40], where it has been used to measure textual diversities like the lexical diversity of online populations [4], the hashtag diversity of online activism [17], and the information content diversity of search engine results [43].

## 2.3 Kullback–Leibler divergence

At times we may want an asymmetric measure of how texts differ. For instance, we may want to measure how language evolved with respect to some reference point in the past

[8], or compare the language of one person to that of an entire community [6]. For these cases, we distinguish between a *reference* text and a *comparison* text. If we let $P^{(1)}$ be the relative word frequency distribution of the reference text and $P^{(2)}$ be the distribution of the comparison, then the Kullback–Leibler divergence (KLD), or relative entropy, is defined as:

$$D^{(\mathrm{KL})}\bigl(P^{(2)} \parallel P^{(1)}\bigr) = \sum_{\tau \in \mathcal{T}} p_\tau^{(2)} \log_2 \frac{1}{p_\tau^{(1)}} - p_\tau^{(2)} \log_2 \frac{1}{p_\tau^{(2)}}.$$

The KLD is the average number of extra bits per word required to encode the words of text $T^{(2)}$ using an optimal coding scheme for $T^{(1)}$ instead of $T^{(2)}$. As such, it shares a form similar to entropy where each word's contribution is the difference between the surprisal of the word in the reference and comparison, but, in contrast to entropy, both surprisals are weighted by the word's relative frequency in the comparison text. The KLD is a conceptually useful measure when we have a well-defined vocabulary and a meaningful reference distribution for comparison. However, if there is a single word that appears in the vocabulary of the comparison but not the reference (i.e., $p_\tau^{(2)} > 0$ and $p_\tau^{(1)} = 0$), then the KLD is infinite. This makes the KLD a brittle measure for comparing texts in general because it is only applicable if the comparison text uses a subset of words from the reference text's lexicon, which is very often not the case when comparing two distinct corpora.

### 2.4 Jensen–Shannon divergence

The Jensen–Shannon divergence (JSD) accounts for some of the shortcomings of the Kullback–Leibler divergence. The JSD compares the similarity of the word distributions by first constructing a probability distribution $M$ for some artificial hybrid text:

$$M = \pi_1 P^{(1)} + \pi_2 P^{(2)}. \tag{5}$$

The mixture weights $\pi_1$ and $\pi_2$ must sum to 1 and are often set to be either equal, $\pi_1 = \pi_2 = 1/2$, or proportional to the number of word tokens in $T^{(1)}$ and $T^{(2)}$. The JSD is then computed as the average KLD from the mixture text,

$$D^{(\mathrm{JS})}\bigl(P^{(1)} \parallel P^{(2)}\bigr) = \pi_1 D^{(\mathrm{KL})}\bigl(P^{(1)} \parallel M\bigr) + \pi_2 D^{(\mathrm{KL})}\bigl(P^{(2)} \parallel M\bigr).$$

By construction, the JSD is symmetric and does not infinitely diverge like the KLD because $M$ consists of the entire vocabulary of both texts. Conveniently, the JSD takes on a value of 0 if the texts are identical and a value of 1 if they have no words in common (as long as we are using base 2 logarithms). The individual contribution $\delta\mathrm{JSD}_\tau$ of a word $\tau$ to the JSD is given by,

$$\delta\mathrm{JSD}_\tau = m_\tau \log \frac{1}{m_\tau} - \left(\pi_1 p_\tau^{(1)} \log \frac{1}{p_\tau^{(1)}} + \pi_2 p_\tau^{(2)} \log \frac{1}{p_\tau^{(2)}}\right), \tag{6}$$

the (corpus-weighted) difference between the surprisal of the word in the average text and the average surprisal of the word in each observed text. Note, the contribution is always non-negative, and $\delta\mathrm{JSD}_\tau = 0$ if and only if $p_\tau^{(1)} = p_\tau^{(2)}$. Like Shannon entropy, the JSD can be generalized to emphasize different regions of the word frequency distribution [42] (see

Materials and methods for details). The symmetric nature of the JSD has made it a useful tool for investigating cultural evolution across digitized collections [16], charting fluctuations in the birth and death of words [10], and disentangling viewpoints in online political discussions [17].

### 2.5 Dictionary-based scores

The measures that we have introduced so far all compare texts based on the relative frequencies of their words. The differences between them lie in how they weight each contribution, where those weights are themselves functions of word frequency. Very often though, we have external weights that we want to specify for each word. The most common example of this is dictionary-based sentiment analysis [4, 44, 45], where we have a dictionary of words and each word is assigned a weight or score according to its association with a particular emotion or feeling. Other dictionaries and lexicons have been curated to encode constructs like morality [23], respect [14], profanity [24], and hatefulness [25].

When we are equipped with dictionary scores, we can calculate the average score of each text as a whole and then compare them. If we have a single dictionary that prescribes a score $\phi_\tau$ for each word $\tau$ in the vocabulary $\mathcal{T}$, then the difference between the weighted averages $\Phi^{(1)}$ and $\Phi^{(2)}$ is

$$\delta\Phi = \sum_{\tau\in\mathcal{T}} \phi_\tau \left(p_\tau^{(2)} - p_\tau^{(1)}\right). \tag{7}$$

When the dictionary does not cover the entire vocabulary (as is often the case), we typically subset the vocabulary to only words appearing in the dictionary. Like the other measures, we can use the linearity of the weighted averages to extract the contributions $\delta\Phi_\tau$ to the difference and rank them accordingly.

## 3 Word shift graphs

When using any weighted average for pairwise text comparison, we want to be able to interpret differences between measurements. Each of the measures that we have introduced can be decomposed into word-level contributions, and so we can identify *which* words most account for the between-text variation. We would like to go further and explain *how* each word contributes. Is one set of lyrics happier than another because it uses more positive words or because, instead, it uses less negative words? Does a social bot's language seem unpredictable because it uses a variety of surprising words or because it uses common words in a surprising way? To what extent do misogynistic internet communities not only use sexist slurs, but also associate other words with negative overtones? These are the kinds of qualitative and contextual questions that can be answered quantitatively through the word shift framework and visualized through word shift graphs.

### 3.1 Word shift fundamentals

We first revisit basic word shift graphs which we first introduced in ref. [18] in the context of happiness measurements, and further developed in refs. [4] and [46]. Basic word shifts are for use when we have single set of scores unchanged across texts [4], as is often the case for (but in no way limited to) standard dictionary-based sentiment analyses. We then generalize the word shift framework so that each text can be equipped with its own set of scores for each word. Finally, we describe and present examples of our generalized word shift graphs, showing how they create detailed summaries of how two texts differ.

As we have been doing, let us say that we have two texts $T^{(1)}$ and $T^{(2)}$ with relative word frequency distributions $P^{(1)}$ and $P^{(2)}$. Suppose, for now, that we have a single dictionary which assigns a score $\phi_\tau$ to each word $\tau$ in the vocabulary $\mathcal{T}$. Our main quantity of interest is the difference between the weighted averages $\Phi^{(1)}$ and $\Phi^{(2)}$,

$$\Phi^{(2)} - \Phi^{(1)} = \sum_{\tau \in \mathcal{T}} \phi_\tau p_\tau^{(2)} - \sum_{\tau \in \mathcal{T}} \phi_\tau p_\tau^{(1)}. \tag{8}$$

Denoting the difference as $\delta\Phi$, we can write it as the sum of contributions from each individual word,

$$\delta\Phi = \sum_{\tau \in \mathcal{T}} \phi_\tau \left(p_\tau^{(2)} - p_\tau^{(1)}\right) = \sum_{\tau \in \mathcal{T}} \delta\Phi_\tau, \tag{9}$$

where we have introduced the notation $\delta\Phi_\tau$ for the summand.

To unpack the qualitatively different ways that words can contribute, we introduce $\Phi^{(\text{ref})}$, a *reference score*. Consider the case of sentiment analysis. For each word in our score dictionary, we not only know its score, but also whether it is considered more or less positive. Importantly, the notion of being "more" or "less" positive is relative to some *reference* value. For example, we may consider a word positive or not based on its position in an overall score distribution—words that are above the average score are positive and those that are below are not. Or instead, we may want to know which words make one text more positive than the other, in which case we can treat the average sentiment of one of the texts as the reference score to determine which words are relatively positive. The quantity $\Phi^{(\text{ref})}$ encodes these kinds of reference points, distinguishing between different regimes of interest among word scores.

Using the reference score $\Phi^{(\text{ref})}$, we can equivalently rewrite the sum of contributions (Eq. (9)) as

$$\delta\Phi = \sum_{\tau \in \mathcal{T}} \left(p_\tau^{(2)} - p_\tau^{(1)}\right)\left(\phi_\tau - \Phi^{(\text{ref})}\right), \tag{10}$$

opening up a richer set of textual interpretations. Each word contribution is now the product of two components: the difference between the word score and the reference score, and the difference between relative frequencies. Both components can be either positive or negative, which yields four different ways that a word can contribute,

$$\delta\Phi_\tau = \underbrace{\left(p_\tau^{(2)} - p_\tau^{(1)}\right)}_{\uparrow/\downarrow}\underbrace{\left(\phi_\tau - \Phi^{(\text{ref})}\right)}_{+/-}. \tag{11}$$

If we say that $\phi_\tau > \Phi^{(\text{ref})}$ implies that a score is "relatively positive," and that $\phi_\tau < \Phi^{(\text{ref})}$ implies that a score is "relatively negative," then without loss of generality we can colloquially phrase the ways that $T^{(2)}$ can have a higher score than $T^{(1)}$ as follows:

1. A relatively positive word (+) is used more often (↑) in $T^{(2)}$ than in $T^{(1)}$.
2. A relatively negative word (–) is used less often (↓) in $T^{(2)}$ than in $T^{(1)}$.

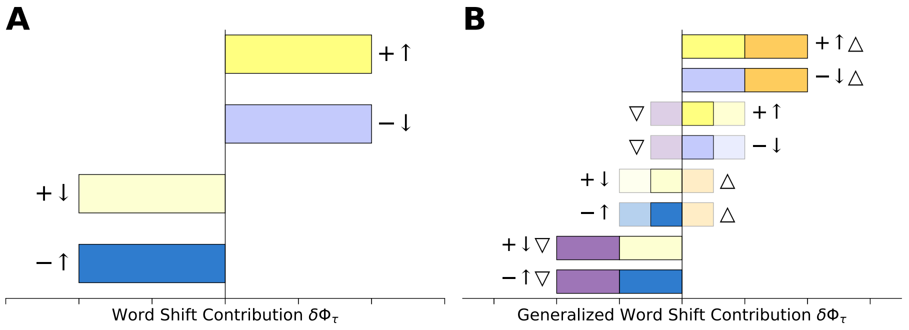
A
+↑
−↓
+↓
−↑
Word Shift Contribution $\delta\Phi_\tau$
B
+↑△
−↓△
▽ +↑
▽ −↓
+↓ △
−↑ △
+↓▽
−↑▽
Generalized Word Shift Contribution $\delta\Phi_\tau$

**Figure 1** Types of word contributions in word shift graphs. (**A**) Word contributions in basic word shift graphs, which are determined by the interaction between the signs of the difference between the word score and the reference score (+/–) and the difference in relative frequencies (↑ / ↓) (see Sect. 3.1). For example, in sentiment analysis, a relatively positive word appearing more is indicated by a deep yellow bar to the right (+ ↑), while a relatively negative word appearing more is indicated by a deep blue bar to the left (– ↑). (**B**) Word contributions in generalized word shift graphs, which additionally visualize the difference in word score (△/▽) (see Sect. 3.2). If component contributions counteract one another then they are faded to emphasize the magnitude of the resulting contribution while retaining information about the detraction of one component from the other. For example, in sentiment analysis, if a relatively positive word is used more and its score is higher across contexts, then it is indicated by a deep yellow bar with an adjacent orange bar, both directed to the right (+ ↑ △). If it is a relatively positive word that is used more but its score is lower, then the components counteract one another, indicated by a deep yellow bar to the right faded by the same amount as a subtractive purple bar to the left (+ ↑ ▽)

Similarly, if $T^{(2)}$ has a higher score than $T^{(1)}$, two types of contributions counteract it to give $T^{(2)}$ a *lower* score than it would have otherwise:

1. A relatively positive word (+) is used less often (↓) in $T^{(2)}$ than in $T^{(1)}$.
2. A relatively negative word (–) is used more often (↑) in $T^{(2)}$ than in $T^{(1)}$.

While the language of "positive" and "negative" most conveniently maps onto the case of sentiment analysis, it is easily altered for other measures, e.g. a word may be "relatively angry" or "relatively surprising" if its score is larger than the reference score.

These contributions are the visual building blocks of word shift graphs (see Fig. 1A). If a word contribution is positive, $\delta\Phi_\tau > 0$ (i.e., + ↑ or – ↓), then the bar points to the right, and if it is negative, $\delta\Phi_\tau < 0$ (i.e. + ↓ or – ↑), then it points to the left. We use color and shading to differentiate the two different ways that each word can contribute in either direction. Relatively positive words (+) are colored in yellow and relatively negative words (–) are colored in blue, which is intuitive for sentiment word shifts, and colorblind friendly for any shift graph in general. Contributions that are due to an increase in word frequency (↑) are shaded with deeper yellows and blues, while contributions from a decrease in word frequency (↓) are shaded with lighter variations of the same colors. The direction, color, and shading succinctly summarize the four qualitatively different ways a word can contribute to the measurement variation between two texts.

### 3.2 Generalized word shifts

Already, we can start to see the richness that word shifts reveal. However, we also want to be able to account for words that have different scores in each corpus, such as with any of the entropy-based measures we introduced, or in sentiment analysis using domain-adapted score dictionaries [35].

We introduce generalized word shifts, which allow words to take on corpus-specific weights. Rather than specifying a single score $\phi_\tau$ across both texts, let $\phi_\tau^{(i)}$ indicate that a word's score can be dependent on its appearance in either $T^{(1)}$ or $T^{(2)}$. The difference in weighted averages $\delta\Phi$ can then be written as

$$\delta\Phi = \sum_{\tau\in\mathcal{T}} \left(p_\tau^{(2)} - p_\tau^{(1)}\right)\left[\frac{1}{2}\left(\phi_\tau^{(2)} + \phi_\tau^{(1)}\right) - \Phi^{(\text{ref})}\right] + \frac{1}{2}\left(p_\tau^{(2)} + p_\tau^{(1)}\right)\left(\phi_\tau^{(2)} - \phi_\tau^{(1)}\right), \tag{12}$$

where we provide full details of the derivation in the Materials and methods. If the scores are the same, $\phi_\tau^{(1)} = \phi_\tau^{(2)}$, then we recover the basic word shift. When the word scores are, in fact, different, the average score of $\phi^{(1)}$ and $\phi^{(2)}$ is compared to the reference $\Phi^{(\text{ref})}$ to determine if the word is "relatively positive" or "relatively negative." The second, new component in the generalized word shift accounts for the difference between the scores themselves, and weights it by the average frequency of the word. So in the generalized word shift framework, there are three major components to how a word contributes,

$$\delta\Phi_\tau = \overbrace{\left(p_\tau^{(2)} - p_\tau^{(1)}\right)}^{\uparrow/\downarrow}\overbrace{\left[\frac{1}{2}\left(\phi_\tau^{(2)} + \phi_\tau^{(1)}\right) - \Phi^{(\text{ref})}\right]}^{+/-} + \underbrace{\frac{1}{2}\left(p_\tau^{(2)} + p_\tau^{(1)}\right)\left(\phi_\tau^{(2)} - \phi_\tau^{(1)}\right)}_{\triangle/\triangledown}. \tag{13}$$

This gives us eight distinct ways that a word can be visualized in a word shift graph (see Fig. 1B). Similar to before, we can visualize the interaction between the difference in relative frequency and the distance from the reference scores as yellow and blue bars. The new component, the difference between scores, is additive, which means that we can visualize it as an additional bar that augments or diminishes the base bar. When the signs of the two components of Eq. (13) are congruent (as in the top two and bottom two bars of Fig. 1B), then we can visualize the score difference as an orange (△) or purple (▽) stacked bar adjacent to the other. However, the two components can also counteract one another. In this case, each component's bar falls in a different direction, and we highlight this tension by coloring the contribution that remains after the counteraction, and fading the underlying offsetting components accordingly (as in the middle four bars). This maintains the full information about a particular word's components while emphasizing the word's overall contribution. The purple, orange, yellow, and blue bars are mutually colorblind friendly.

### 3.3 Generalized word shift graphs

We now present generalized word shift graphs in their entirety. For our visual case study, we compare the average sentiment of speeches by two United States presidents: Lyndon B. Johnson (1963–1969) and George W. Bush (2001–2009). We use the labMT sentiment dictionary [4] to construct a basic word shift, and the SocialSent historical sentiment lexicons [35] for a generalized word shift. The SocialSent lexicons are decade-specific sentiment dictionaries that were adapted for each decade between 1850 and 2000 by applying semi-supervised machine learning to the Corpus of Historical American English, and so words can take on different scores depending on what sentiment they were associated with in the 1960s or 2000s. We use the word shift graphs (presented in Fig. 2) primarily as visual examples, and so we focus more on their construction and layout rather than their substantive interpretation.

We measure the difference in average sentiment of the presidential speeches, $\Phi^{(\text{G.W.B.})} - \Phi^{(\text{L.B.J.})}$, and rank words by their absolute contribution to that difference. According to both

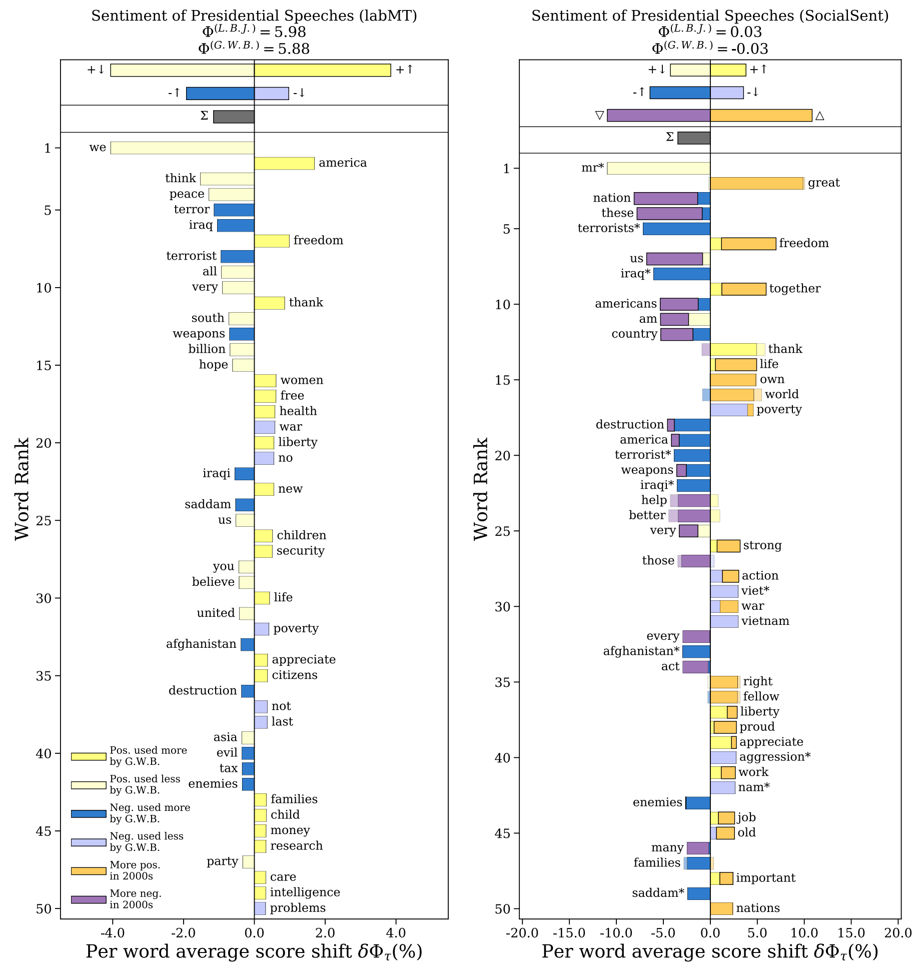

**Figure 2** Word shift graphs of the sentiment of presidential speeches by United States Presidents Lyndon B. Johnson and George W. Bush. We display a basic word shift graph on the left, using the labMT sentiment dictionary [4] (one score per word); on the right is a generalized word shift graph, using the SocialSent decade-adapted sentiment dictionaries [35] (one score per word per decade). Word shifts show the top fifty contributing words to the difference in sentiment. Words to the left are those that contribute to Bush's speeches being more negative than Johnson's, while words to the right partly offset that negativity. Bars at the top show the overall sentiment difference and the effect of each type of word contribution on that difference. The bottom left corner includes a legend for interpreting the different types of contributions in the context of the presidential speeches. In the generalized word shift graph, words that are borrowing a score across decades are marked by an asterisk (∗). We use the center of each of labMT's and SocialSent's sentiment distributions as the reference scores: $\Phi^{(\mathrm{ref})} = 5$ for labMT and $\Phi^{(\mathrm{ref})} = 0$ for SocialSent

dictionaries that we have employed, Bush's speeches were more negative than Johnson's, as indicated by the average sentiments displayed in the title of each graph. We plot word contributions as a horizontal bar chart, where words that contribute to the negativity of Bush's speeches are directed to the left, while words that counteract $\Phi^{(\mathrm{G.W.B.})} < \Phi^{(\mathrm{L.B.J.})}$ point to the right. We provide a legend for the qualitative interpretation of these bars in the bottom left corner of Fig. 2.

Examining the basic word shift graph on the left, we see that Bush's use of more negative words (– ↑), like 'terror', 'weapons', and 'tax', all lower the sentiment of his speeches relative to Johnson. Further, the decreased use of positive words (+ ↓), such as 'we', 'peace', and 'hope', also contributes to the negativity of Bush's speech. On the other hand, these contributions are partly offset by a lesser use of negative words ($- \downarrow$) like 'no', 'poverty', and 'problems', and greater use of positive words ($+ \uparrow$) like 'america', and 'freedom'.

In the generalized word shift to the right, we see that changes in the sentiments of the words themselves also affect the overall difference between Bush's and Johnson's speeches. The words 'nation', 'us', and 'destruction' are all associated with more negativity ($\bigtriangledown$) in the 2000s than in the 1960s. Similarly, but in the opposite direction, 'freedom', 'together', and 'life', are all associated with more positivity ($\bigtriangleup$) in the 2000s than the 1920s. We also see counteracting contributions for individual words: 'better', for example, is a positive word that was used more by Bush, but its positive contribution is offset by its decline in sentiment from the 1960s to the 2000s.

At the top of both figures, we display how each distinct type of word shift contributes to the total difference, $\Sigma$. In the basic word shift graph, we see that the negativity of Bush's speeches is most explained overall by the use of more negative words ($- \uparrow$) and less positive words ($+ \downarrow$). In the generalized word shift, the negativity is most affected by the general negative shift in word sentiment ($\bigtriangledown$) from the 1960s to the 2000s, though that component is largely offset by other words increasing in sentiment ($\bigtriangleup$). These summary totals help accumulate sentiment information across all of the words and tell us what qualitative factors play the largest roles in differentiating the speeches of Presidents Bush and Johnson.

Note that unlike when using a single sentiment dictionary, the SocialSent historical sentiment lexicons have overlapping but distinct vocabularies. For example, because a war with Iraq was not a major topic of discussion in the 1960s, the word 'iraqi' does not appear in SocialSent's 1960s lexicon. Similarly, the word 'viet' does not appear in SocialSent's 2000s lexicon. However, in the 1960s, the word 'viet' carries important emotional information, as does 'iraqi' in the 2000s. To retain some of this information without discarding it entirely, we "borrow" scores across decades that exist in one sentiment lexicon but not another. That is, for example, we use the 1960s sentiment score for 'viet' when considering its appearances in Bush's speeches made in the 2000s. We append an asterisk ($*$) to any word that borrows a score. The borrowing nullifies the component of the word shift contribution that quantifies the change in the score, but it allows us maintain information about the change in frequency and the relative positivity or negativity of a word. In the Materials and methods, we discuss in more detail when "borrowing" may or may not be appropriate.

We also highlight the example of 'iraqi' as a point of caution. In both the labMT and SocialSent dictionaries, the words 'iraqi', 'iraq', and 'afghanistan' are "negative" words whose usage make Bush's speeches quantitatively more negative than Johnson's. However, it is important to keep in mind how these scores were derived. For labMT, the scores are a product human annotations collected in the late 2000s when the United States' wars in the Middle East were ongoing. For SocialSent, the scores are a product of machine learning, where commonly co-occurring words share similar sentiment, and so the negativity is likely a product of those words appearing frequently with other words like 'war' and 'terrorist'. Similarly, we would likely find that the word 'woman' is associated with negative sentiment if we made dictionaries for misogynistic internet forums, and that the phrase 'African American' is "negative" if we made dictionaries for pamphlets produced by white supremacists. In all of these cases, the negative quantitative scores should *not* be interpreted as indicating that these people and countries are bad, unworthy of respect, or deserving of oppression. The scores reflect biases of the cultural contexts in which they were produced and should be interpreted accordingly. Fortunately, the word shift graphs make these biases clear, rather than burying them under an aggregate measure. However, while word shift graphs are interpretable, we still need to take care with the actual interpretation itself so that we do not reproduce systemic inequalities in our own analyses.

Overall, the generalized word shift graph succinctly visualizes which words contribute to the negativity of George W. Bush's speeches relative to Lyndon B. Johnson and, importantly, how they do so. The word shift graphs distinguish between subtle differences in contributions, such as whether the speeches are more negative because more negative words were used or less positive ones were. Rather than just comparing two averages, like $\Phi^{(\text{G.W.B.})} = -0.03$ and $\Phi^{(\text{L.B.J.})} = 0.03$, the word shift graphs allow us to simultaneously quantify word usage, sentiment, bias, and temporal drift to tell a richer story about how, plausibly, Bush's speeches were negative in part due to their focus on the Iraq War starting in 2003 and, perhaps, also in part due to decreased positivity associated with nationalistic words like 'nation', 'us', 'country', 'america', and 'americans'.

### 3.4 Pairwise comparison measures as word shifts

We have shown how dictionary scores can be naturally incorporated into the word shift framework. We now return to the other text comparison measures that we introduced earlier: relative frequency, Shannon entropy, the Kullback–Leibler divergence (KLD), the Jensen–Shannon divergence (JSD), and their generalized forms (see Materials and methods for details). Some of these measures, like relative frequency and the Shannon entropy, are easily identifiable as weighted averages by how they are commonly written. Other measures though, like the KLD, the JSD, and the generalized entropies, can often be expressed in ways that do not make it clear that they are also weighted averages. In Table 1, we explicitly write the word contribution $\delta\Phi_\tau$ of each measure as a difference in weighted averages. Making this form explicit allows us to easily situate all of these measures within the generalized word shift framework and visualize them through generalized word shift graphs.

Formulating each measure in terms of weighted averages is one of two key elements for using generalized word shift graphs. The other is identifying a reference score $\Phi^{(\text{ref})}$ that discerns between distinct and interesting regimes of the word scores. As we have seen with sentiment analysis, one obvious candidate for the reference score is the center of the sentiment scale, which naturally sifts positive words from negative ones. While, in practice, researchers rarely draw an explicit boundary between different types of words when

**Table 1** Contributions and scores of various text comparison measures according to the word shift framework. The word contribution $\delta\Phi_\tau$ indicates how an individual word impacts a measure, and each contribution is expressed as a difference in weighted averages so that it can be easily identified with the components of the word shift framework

| Measure | Notation | Word contribution $\delta\Phi_\tau = p_\tau^{(2)}\phi_\tau^{(2)} - p_\tau^{(1)}\phi_\tau^{(1)}$ |
|---|---|---|
| Relative Frequency | $P^{(i)}$ | $p_\tau^{(2)} - p_\tau^{(1)}$ |
| Shannon Entropy | $H(P^{(i)})$ | $-p_\tau^{(2)} \log p_\tau^{(2)} + p_\tau^{(1)} \log p_\tau^{(1)}$ |
| Generalized Entropy | $H_\alpha(P^{(i)})$ | $-p_\tau^{(2)}[\frac{(p_\tau^{(2)})^{\alpha-1}}{\alpha-1}] + p_\tau^{(1)}[\frac{(p_\tau^{(1)})^{\alpha-1}}{\alpha-1}]$ |
| Kullback–Leibler Divergence | $D^{(\text{KL})}(P^{(2)} \parallel P^{(1)})$ | $-p_\tau^{(2)} \log p_\tau^{(1)} + p_\tau^{(2)} \log p_\tau^{(2)}$ |
| Jensen–Shannon Divergence | $D^{(\text{JS})}(P^{(1)} \parallel P^{(2)})$ | $p_\tau^{(2)}\pi_2 \log \frac{p_\tau^{(2)}}{m_\tau} - p_\tau^{(1)}\pi_1 \log \frac{m_\tau}{p_\tau^{(1)}}$ |
| Generalized Jensen–Shannon Divergence | $D_\alpha^{(\text{JS})}(P^{(1)} \parallel P^{(2)})$ | $p_\tau^{(2)}\pi_2[\frac{(p_\tau^{(2)})^{\alpha-1} - m_\tau^{\alpha-1}}{\alpha-1}] - p_\tau^{(1)}\pi_1[\frac{m_\tau^{\alpha-1} - (p_\tau^{(1)})^{\alpha-1}}{\alpha-1}]$ |

using the measures presented in Table 1, the generalized word shift framework provides us an opportunity to be more intentional and creative with how we quantify, interpret, and visualize differences between texts. For example, researchers often remove commonly appearing "stop words" [3, 47] by applying a pre-assembled list or identifying the top, say, 1% of frequently occurring words. Rather than discarding them in an ad hoc manner though, we may instead choose to leave them in and use them to help us mark the boundary between frequent and infrequent, surprising and unsurprising. Or if we are working with an entropy-based measure, we may set the reference value to be the average entropy of one of the texts. Using the text's entropy as a reference allows us to discern which words contribute to a text's unpredictability because they are even more surprising than the average surprisal.

Of course, it is always mathematically valid to set $\Phi^{(\text{ref})} = 0$ if we are satisfied with just knowing *which* words distinguish two texts. However, doing so always risks masking the richness in *how* those words contribute. Placing the frequency and entropy-based measures within the generalized word shift framework gives us new ways of understanding them and disentangling the complexities of text as data.

## 4 Case studies: using word shift graphs in practice

To show how generalized word shift graphs can be used in practice, we present a diverse set of case studies that highlight how they can be used as both a diagnostic tool during the research process and an illustrative instrument for scientific communication. First, through sentiment analyses of both the book *Moby Dick* and U.S. urban parks, we demonstrate how word shifts warn us when there are significant measurement issues that require us to revisit how the text is preprocessed and quantified. Second, through a case study of Twitter's change from 140 to 280 character tweets, we show how word shifts make it possible to interpret unexplained textual trends and generate additional research hypotheses. Finally, through a case study of labor diversity and the Great Recession, we show how shift graphs enrich analyses beyond just the research process and provide fine-grained evidence that support deeper substantive insights by domain experts.

### 4.1 Sentiment peculiarities of Moby Dick and U.S. urban parks

Dictionary-based sentiment analysis is sensitive, of course, to the dictionary that is used. Sentiment dictionaries are often static objects, constructed once for general use. This can be problematic if there has been a temporal shift in how particular words or used, or when words take on different sentiments in particular contexts [3, 26]. As we show, word shift graphs transparently diagnose these kinds of measurement issues.

We start with a case study of *Moby Dick*, the 1851 novel by Herman Melville. We naively apply the labMT sentiment dictionary [4] to the first and second halves of the book, a simple quantification of the novel's emotional arc [19]. The sentiment word shift graph is shown on the left in Fig. 3. There are two issues that are made visible by the word shift graph, each of which we could easily miss otherwise. First, examining the left panel of Fig. 3, the overall sentiment is affected considerably by the words 'cried' and 'cry.' Throughout the book though, 'cried' and 'cry' are often understood to mean 'said.' Second, the word 'coffin' also significantly affects the sentiment. However, while coffins are mentioned throughout the story, searching the raw text of the novel[a] reveals that about a third of its usage is with respect to the surname 'Coffin.' All three of these words contribute in an unintended way to the sentiment differences between the first and second half the book—with

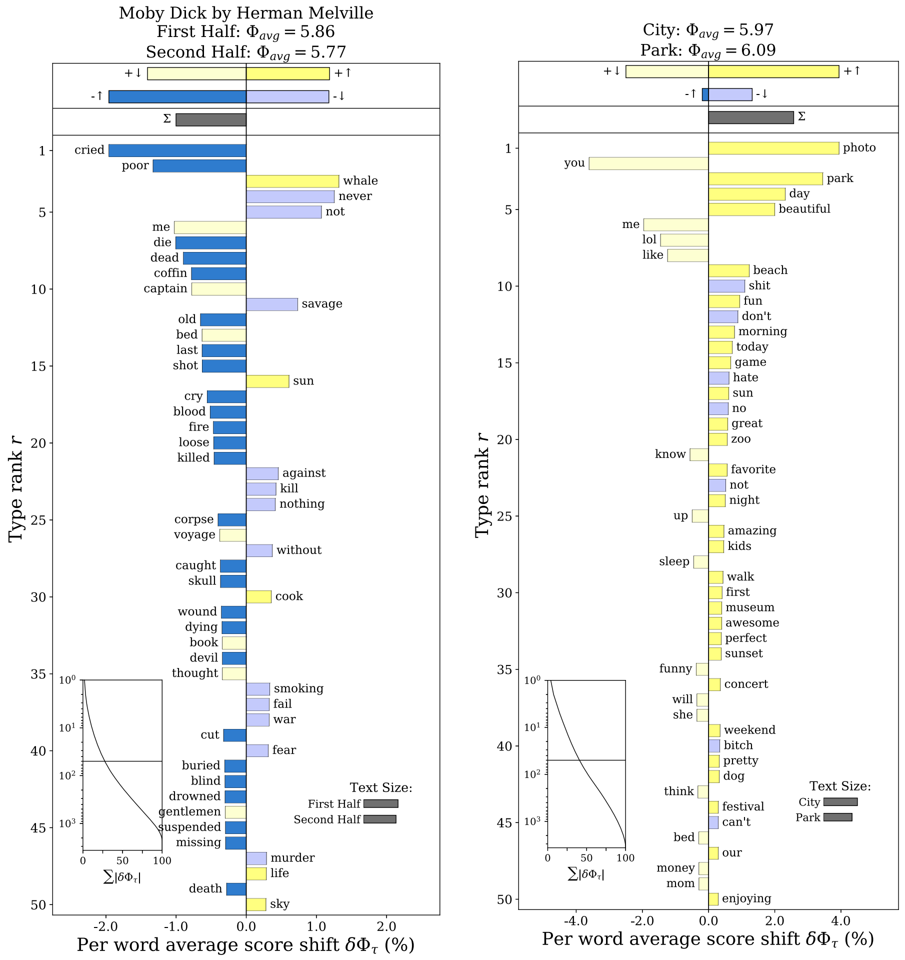

**Figure 3** (*Left*) Word shift graph of the sentiment difference between the first and second halves of *Moby Dick* by Herman Melville. A naive application of a dictionary-based sentiment lexicon for the two-segment emotional arc would inflate the negative trajectory of the novel without the preprocessing or removal of words like 'cried' and 'cry', which more often means 'said' in nineteenth century English, and 'coffin', which is used as a surname about one third of the time. We find that $\delta\Phi = 0.09$ when including those words, while $\delta\Phi = 0.07$ when they are excluded. (*Right*) Word shift graph of the sentiment difference between the in-park and out-of-park tweets across 25 cities in the US. A naive application of a dictionary-based sentiment lexicon would inflate the in-park tweet scores by including words like 'park', 'beach', 'zoo', 'museum', 'music', and 'festival', all of which represent physical locations and events within parks. We find that $\delta\Phi = 0.12$ when including those words, while $\delta\Phi = 0.10$ when they are excluded. For both word shift graphs, a reference value of 5 was used, and a stop lens was applied on all words with a sentiment score between 4 and 6. Both word shift graphs contain cumulative contribution and text size diagnostic plots in their bottom left and right corners, respectively. See the following case study and the Materials and methods for more details on their interpretation

them included, the difference in sentiment is $\delta\Phi = 0.09$, while without them it is $\delta\Phi = 0.07$, a difference of over 20%.

Word shift graphs make these contributions apparent. One way to address these issues is through additional text preprocessing. For example, removing only capitalized uses of 'Coffin' (along with 'cry' and 'cried') allows for 'coffin' to still contribute, yielding a sentiment difference of $\delta\Phi = 0.08$, which is 15% less than the naive approach. Another way is through modification of the dictionary itself—domain knowledge or semi-supervised machine learning [35] can help refine or adapt the sentiment dictionary to the language of nineteenth century English. By highlighting these mismeasurements early in the research process, word shift graphs allow researchers to make appropriate adjustments in the data pipeline.

To emphasize the need for word shift graphs in identifying bias induced by sentiment mismeasurement, we also consider a case study of tweets posted inside and outside of U.S. urban parks. Prior work has demonstrated that people are happier when visiting urban green spaces such as parks [48], and social media data presents an opportunity to supplement traditional survey measures with geographically fine-grained measurements. However, naively applying the labMT sentiment dictionary to tweets may overestimate the sentiment difference between in- and out-of-park tweets. In the right panel of Fig. 3, we see that the word 'park' is contributing substantially to the higher sentiment of in-park tweets. However, in the context of inferring happiness from tweets, writing the word 'park' is often simply a declaration of where a user is located, rather than a proxy for how they may be feeling. Similarly, words like 'museum', 'zoo', and 'beach' also represent physical locations within parks, but contribute to the positivity of in-park tweets because they are all relatively positive words. 'Music' and 'festival' also appear frequently within park tweets, which are related to events in parks, but often not nature itself.

While there are defensible arguments for and against removing each of these words, word shift graphs make their contributions visible, and allow a researcher to make transparent decisions with the understanding of how results may change based on which words are included in the final analysis. When removing the above six words, the sentiment difference goes from $\delta\Phi = 0.12$ to $\delta\Phi = 0.10$, more than a 15% difference. Adjustments for specific words, in tandem with the examination of a word shift graph, allow us to apply sentiment analysis with the confidence that one or a few individual words have not made a folly of our analyses.

### 4.2 Information content of 280 character tweets

On November 7th, 2017, Twitter doubled the character limit for all tweets from 140 to 280 characters, one of the most significant changes to the platform since its inception in 2006. Prior to the change, Twitter found that there were discrepancies in how often users reached the 140 character limit based on the language in which they tweeted [49]. They partly attributed the discrepancies to the ability of different languages to encode more or less information in a single character [49, 50]; for example, users tweeting in English hit the limit 9% of the time, while those tweeting in Japanese rarely did so. Immediately following the update to 280 character tweets, English users only hit the limit about 1% of the time, suggesting that the change made it "easier to tweet" by making it easier for individuals to spend "less time editing their tweets in the composer" [51]. Outside of Twitter, it has been independently verified that the increased character limit increased the Flesch-Kincaid reading level of tweets and decreased the proportion of users hitting the character limit [52].

Moving from individual characters to whole words, we measure the Shannon entropy of the Twitter word distribution before and after the 280 character change to understand how the character-level change may have affected the information content of the language used in tweets. We take all tweets collected from Twitter's Decahose, a 10% sample of tweets, 30 days before and after the change and aggregate them into two separate bags of words, using the labMT dictionary as our vocabulary (see Materials and methods for

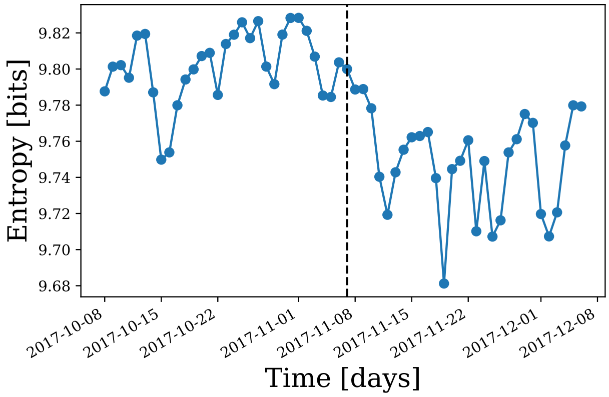

**Figure 4** Time series indicating the change in the entropy of Twitter language before and after the platform's change from 140 character to 280 character tweets (marked by dashed line)

details). Figure 4 shows that the information content of tweet language decreased after the change to 280 characters. Using a generalized word shift graph, we can reveal what words specifically contributed to that drop and why (see Fig. 5). We use the entropy before the change to 280 character tweets as our reference value, implying that a word is considered relatively "surprising" if its surprisal is higher than the average word surprisal in 140 character tweets.

As we may expect from a change allowing for longer tweets, the top five contributions to the decrease in entropy all come from greater use of the most common parts of speech, including conjunctions ('and', 'that'), articles ('the') and prepositions ('to', 'of'). All of these are relatively "unsurprising" words, so they drive the entropy down in 280 character tweets. It is plausible that in the shift to longer tweets, users are able to write longer messages and use less abbreviations, allowing them to place a heavier on traditional function words. Further, the entropy word shift also reveals some more unexpected trends, namely a decrease in first- and second-person personal pronouns ('i', 'you', 'me', 'my', 'u', 'i'm' and 'your') and an increase in third-person pronouns ('we', 'they', 'their', 'our', and 'them'). This is somewhat striking, particularly as it is an observation that has emerged from the data in an unsupervised manner.

Finally, we note that we have appended an inset plot to the bottom left of the word shift graph. This inset describes the percent of all variation in entropy that is explained by the word shift. It does so by plotting how the difference $\delta\Phi$ cumulatively changes as we successively add word contributions according to their rank. The horizontal line demarcates the boundary between the top forty words shown in the plot and the thousands of other words used in the tweets. We use this plot in conjunction with the observation that the majority of the top forty contributions are from relatively unsurprising words. As shown by where the cumulative curve intersects the horizontal cutoff line, these explain a bit more than 30% of the total entropy difference between 140 character language and 280 character language. Yet, by the top of the word shift graph, we see that the largest contributions come from the use of relatively surprising words, few of which appear in our figure. This suggests that there is a richer story in the long tail of the word distribution than is shown solely by the word shift graph. This would not be obvious without the cumulative contribution diagnostic plot, which we describe further in the Materials and methods.

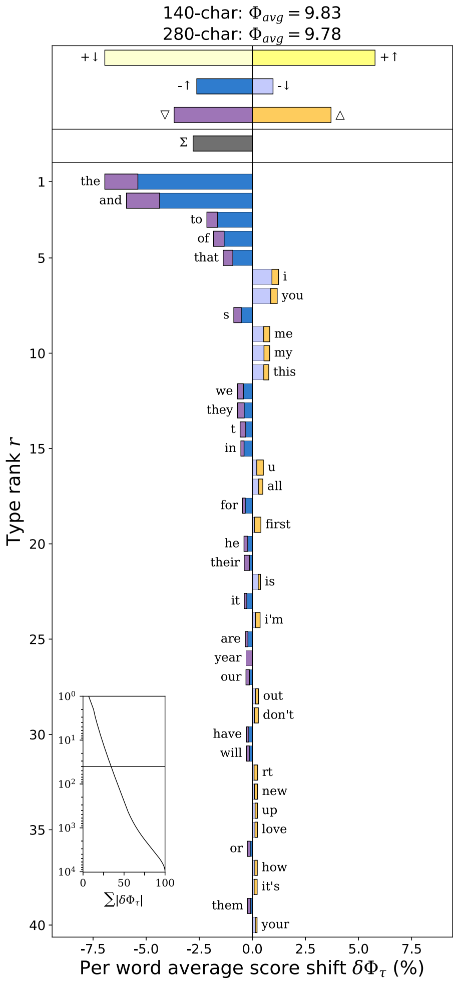

**Figure 5** Generalized word shift graph of the change in Shannon entropy for the 30 days before and after the 140 to 280 character change on Twitter. Words are relatively "surprising" (+) or "unsurprising" (–) depending on if their surprisal is higher than the entropy, or average surprisal, of words used in 140 character tweets. Many of the contributions in the top forty words are from unsurprising words being used relatively more (– ↑) or less (– ↓). The surprisal for each word went down (▽) or up (△) depending on if it was used more or less, respectively, in 280 character tweets. The inset in the bottom left shows how the word shift scores cumulatively vary as a function of word rank, where the horizontal line in the middle of the plot indicates what is explained by the top forty words shown in the graph (see Materials and methods for more details). As seen by the top of the word shift graph and the cumulative inset, most contributions come from a long tail of relatively surprising words, despite mostly not appearing among the top forty words

Through a brief investigation of the change from 140 to 280 character tweets, generalized word shift graphs have allowed us to uncover three potentially fruitful hypotheses: Twitter users do not need to abbreviate common function words as often, tweets deploy more collective framing through third-person pronouns, and less common words account for the largest shift in entropy. Of course, all of these hypotheses are speculative and require much deeper investigations. This is exactly what demonstrates the power of word shift graphs though. These stories are hidden by the aggregate entropy measures, which obscure why the entropy dropped after the character limit change. Generalized word shift graphs unpack these measures and allow us to quickly generate new questions and hypotheses that bring our research in directions that may have been otherwise unexplored.

### 4.3 Employment diversity and urban resilience during the Great Recession

The Great Recession, which spanned from the end of 2007 to 2012, is one of the most significant economic disruptions in the United States' history [53]. Understanding which U.S. cities were more or less resilient to the recession and why could inform urban policy that diminishes the disruptions to labor and employment. Similar to ecological systems [54], and complex systems more generally, employment diversity is hypothesized to play a key role in the resilience of urban labor markets. Diverse labor markets have more potential for redundancies among occupations that enable local workers and firms to adapt to disruptions. For example, one recent study compared labor and skill diversity in cities to the cities' exposure to automation [55] and found that labor market diversity was more predictive than the size or regional economy of the city.

To study urban response to the recession, we turn to the U.S. Bureau of Labor Statistics (BLS), which records employment data for cities across the United States. If we consider each city to be a "corpus" and each distinct occupation to be a "word," then we can use the word shift framework to understand differences in employment diversity between cities. Let $\mathcal{J}$ be the "vocabulary" of the 794 jobs recorded by the BLS in 2007 across 375 U.S. cities, and denote the number of people employed with job $j$ in city $c$ as $f_j^{(c)}$. The total employment across the entire urban labor market is $T^{(c)} = \sum_{j \in \mathcal{J}^{(c)}} f_j^{(c)}$, and the relative frequency of a job in the labor distribution $P^{(c)}$ is $p_j^{(c)} = f_j^{(c)} / T^{(c)}$, like our word distributions.

Traditionally, many labor economists consider job markets to be "deep" or "shallow" depending on the number of unique occupations, which we refer to as *occupation diversity*, $N^{(c)} = |\mathcal{J}^{(c)}|$. We can more fully account for the distribution of worker employment by considering the *employment diversity*, the Shannon entropy [40, 41] of the worker employment distribution,

$$H\left(P^{(c)}\right) = -\sum_{j \in \mathcal{J}} p_j^{(c)} \log_2 p_j^{(c)}. \tag{14}$$

Using these measures, we examine the relationship between labor diversity and peak city unemployment during the Great Recession, as given by the BLS's Local Area Unemployment Statistics, and present the results in Fig. 6 in the form of regression coefficients. As we would expect, total employment is significantly positively associated with raw unemployment counts. After controlling for the total employment though, we find evidence that the occupation diversity $N^{(c)}$ is not significantly associated with unemployment, while employment diversity $H^{(c)}$ is significantly negatively associated with peak unemployment during the Great Recession. The inset in Fig. 6 visualizes this negative association when controlling for total employment and occupation diversity.

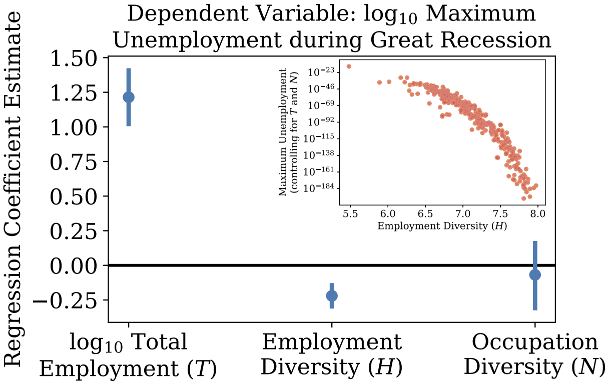

**Figure 6** Linear regression coefficients showing that U.S. cities with greater employment diversity $H^{(c)}$ saw lower peak unemployment during the Great Recession, after controlling for total employment $T^{(c)}$ and the number of unique occupations in each city $N^{(c)}$ ($R^2 = 0.915$). All variables were centered and standardized prior to regression. Error bars represent 95% confidence intervals. Additional details on the regression analysis are provided in the Materials and methods

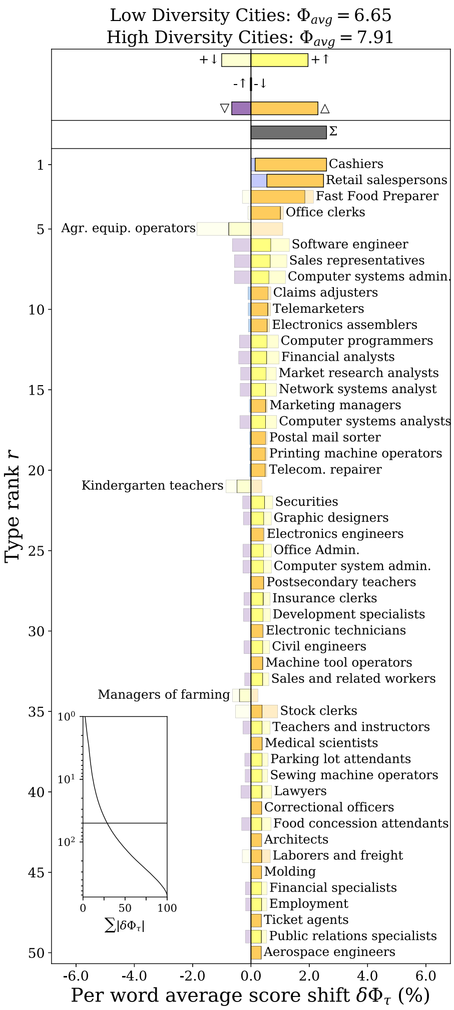

**Figure 7** By treating the employment distribution in each U.S. city as a "text" and each occupation as a "word," we employ a generalized word shift graph to compare the differences in employment between the 15 cities with the most diverse employment distributions and the 15 cities with the least diverse employment distributions, as measured by the Shannon entropy. We use the average employment diversity across cities' workforces as the reference value ($\Phi^{(\text{ref})}$). Occupations are relatively "surprising" (+) or "unsurprising" (–) depending on their surprisal in each class of city

The relationship between lower unemployment and employment diversity $H^{(c)}$, rather than the number of unique jobs $N^{(c)}$, suggests that urban policy may want to focus on growing the employment diversity of a city's workforce to bolster its economic resilience. However, it is not clear from the aggregate employment diversity which occupations are most likely worth the time, money, and effort of designing policy that reshapes the labor market. By using a (word) shift graph, we can quantify the jobs that most distinguished employment differences between the 15 most diverse cities and the 15 least diverse cities in 2007 (see Fig. 7). We consider a job to be relatively more "surprising" or "unsurprising" compared to the entropy of employment distributions averaged across all U.S. cities, i.e. $\langle H^{(c)} \rangle$. As shown by the shift graph, the differences in employment diversity come from two main sources: less common occupations ($+\uparrow$) that are relatively abundant in high diversity cities, and jobs that are common in low diversity cities but less so in high diversity cities ($\triangle$). There are some deviations from these trends; for example, kindergarten teachers and agricultural workers have high surprisal but were relatively more abundant in cities with low employment diversity ($+\downarrow$).

It is beyond the scope of this brief case study to suggest causal policy interventions based on this one shift graph. However, given a more comprehensive study examining labor diversity, unemployment, and economic resilience, it is clear how a shift graph could provide targeted and actionable insights that are not otherwise possible through aggregate measures. The details provided by generalized shift graphs allow us to more deeply draw on our domain knowledge and build a more comprehensive understanding of the social scientific phenomena under study.

## 5 Discussion

We have introduced generalized word shifts, a framework for taking pairwise comparisons between texts and understanding their differences at the word level. In this framework, comparison measures are decomposed into their individual word contributions so that the words can be ranked and categorized according to how they contribute. The word shift form that we have presented generalizes a previous iteration [4, 37], which was limited to single dictionary-based weighted averages. Our generalization naturally incorporates multi-dictionary scoring, the Shannon entropy, generalized entropies, the Kullback–Leibler divergence, the Jensen–Shannon divergence, and any other measure that can be rewritten as a weighted average or difference in weighted averages. All of these generalized word shifts can be summarily visualized as horizontal stacked bar charts, and we have detailed how to effectively interpret the various interacting components of generalized word shift graphs. To help facilitate their use in computational text analyses, we have implemented generalized word shift graphs in an accessible open source Python package, available at https://github.com/ryanjgallagher/shifterator.

Generalized word shift graphs are an interpretative tool that allows researchers to fully harness textual measures, both for their audiences and for themselves. While researchers are often limited to arguing in terms of aggregate weighted averages, generalized word shift graphs provide a principled way of decomposing them into word level characterizations that highlight the most salient differences between two texts. In the best case scenario, when the micro word dynamics exposed by a word shift graph align with a macro research story, visualizing word shifts helps audiences better understand and trust what is being measured. However, generalized word shift graphs are not just visual embellishments to persuade audiences. They are also a robustness check that allow us to convince *ourselves* that we have constructed scientifically sound stories. During the research process, generalized word shift graphs can alert us to data peculiarities, counterintuitive phenomena, and measurement errors. Using generalized word shift graphs as a diagnostic tool gives us the opportunity to catch these oddities, account for them, and better understand our text data. Generalized word shift graphs are immediately applicable to a wide range of computational text analyses, and making them a regular part of the text-as-data workflow promises to enrich the work of many computational social scientists, digital humanists, and other practitioners.

Of course, not every text comparison measure can be formulated in terms of weighted averages. For example, many forms of the commonly used term frequency-inverse document frequency cannot be disentangled into a weighted average. Any non-parametric measure that works with ranks rather than frequencies [56] cannot, by definition, be written as a weighted average. However, while some additive measures like these cannot be retrofitted into the generalized word shift framework that we have outlined here, we still strongly encourage researchers to always visualize the word contributions that differentiate texts, even if just for themselves during exploratory analyses. Linear, additive text comparison measures are inherently interpretable, and we should always make sure to leverage that interpretability to question, improve, and defend the data stories that we discover.

Generalized word shift graphs directly confront the complexity that is inherent in working with text as data. Used together with other methods, tools, and visualization techniques that open up otherwise opaque black-box methods, word shift graphs can help us better triangulate interesting and meaningful social-scientific phenomenon among the vast and ever expanding landscapes of language, stories, and culture encoded in textual data.

## 6 Materials and methods

### 6.1 Generalized entropies

For a relative word frequency distribution $P^{(i)}$, we can calculate its generalized, or Tsallis, entropy of order $\alpha$ [39, 42],

$$H_\alpha = \frac{1}{1-\alpha}\left(\sum_{\tau\in\mathcal{T}} p_\tau^\alpha - 1\right) = \frac{1}{\alpha-1} - \sum_{\tau\in\mathcal{T}} p_\tau^{(i)}\left[\frac{p_\tau^{(\alpha-1)}}{\alpha-1}\right], \tag{15}$$

where the latter form is more recognizable as a weighted average. The parameter $\alpha$ controls how much weight is given to common and uncommon words. When $\alpha > 1$, more weight is given to frequent words. When $\alpha < 1$, more weight is given to rare words. When $\alpha = 1$, we retrieve the Shannon entropy $H_1 = -\sum_\tau p_\tau^{(i)} \log p_\tau^{(i)}$, which marks the information-theoretic boundary between giving preference to frequently or infrequently occurring words [40, 41]. Like the other measures, we can identify a word's contribution by considering the components of $H_\alpha^{(2)} - H_\alpha^{(1)}$,

$$\delta\Phi_\tau = -p_\tau^{(2)}\left[\frac{(p_\tau^{(2)})^{\alpha-1}}{\alpha-1}\right] + p_\tau^{(1)}\left[\frac{(p_\tau^{(1)})^{\alpha-1}}{\alpha-1}\right], \tag{16}$$

where the quantity $1/(\alpha-1)$ cancels out in the difference.

The Jensen–Shannon divergence (JSD) can also be extended through generalized entropies [42]. Recall, for the JSD we form a mixture distribution $M = \pi_1 P^{(1)} + \pi_2 P^{(2)}$, where $\pi_1$ and $\pi_2$ are tunable weights. Rather than writing the JSD as an average KLD relative to the mixture distribution, we can equivalently formulate the JSD in terms of the generalized entropy,

$$D_\alpha^{(\mathrm{JS})} = H_\alpha(M) - \pi_1 H_\alpha\left(P^{(1)}\right) - \pi_2 H_\alpha\left(P^{(2)}\right). \tag{17}$$

With some rearranging, we can write this as a single sum across words and identify their word shift form,

$$\delta\Phi_\tau = p_\tau^{(2)}\pi_2\left[\frac{(p_\tau^{(2)})^{\alpha-1} - m_\tau^{\alpha-1}}{\alpha-1}\right] - p_\tau^{(1)}\pi_1\left[\frac{m_\tau^{\alpha-1} - (p_\tau^{(1)})^{\alpha-1}}{\alpha-1}\right]. \tag{18}$$

Like the generalized entropy, we recover the familiar JSD when $\alpha = 1$. The heavy tail nature of word distributions can make the JSD sensitive to different word frequencies, particularly when we are working with a sample of texts from a larger corpus (which is very often the case) [42]. To obtain more reliable estimates of the JSD for those situations, it is advisable to tune the parameter $\alpha$ acccordingly (see ref. [42] for details).

### 6.2 Derivation of generalized word shifts

Recall, for a word $\tau$ in text $T^{(i)}$, we denote its relative frequency as $p_\tau^{(i)}$ and its (possibly text dependent) score as $\phi_\tau^{(i)}$. The average score across the entire text $T^{(i)}$ is notated as $\Phi^{(i)}$, and the difference in weighted averages is

$$\delta\Phi = \Phi^{(2)} - \Phi^{(1)} = \sum_{\tau \in \mathcal{T}} \phi_\tau^{(2)} p_\tau^{(2)} - \phi_\tau^{(1)} p_\tau^{(1)}. \tag{19}$$

We first introduce $\Phi^{(\mathrm{ref})}$. Note, $\sum_\tau p_\tau^{(i)} = 1$, and so we may write

$$\sum_\tau \Phi^{(\mathrm{ref})} \big(p_\tau^{(2)} - p_\tau^{(1)}\big) = \Phi^{(\mathrm{ref})}(1-1) = 0. \tag{20}$$

Because the entire above quantity is simply zero, we can subtract it from Eq. (19) to get

$$\delta\Phi = \sum_\tau \phi_\tau^{(2)} p_\tau^{(2)} - \phi_\tau^{(1)} p_\tau^{(1)} - \sum_\tau \Phi^{(\mathrm{ref})} \big(p_\tau^{(2)} - p_\tau^{(1)}\big) \tag{21}$$

$$= \sum_\tau p_\tau^{(2)} \big(\phi_\tau^{(2)} - \Phi^{(\mathrm{ref})}\big) - p_\tau^{(1)} \big(\phi_\tau^{(1)} - \Phi^{(\mathrm{ref})}\big). \tag{22}$$

Now, we again use the mathematical sleight of hand of adding zero to rewrite the scores $\phi_\tau^{(i)}$ as

$$\phi_\tau^{(1)} = \frac{1}{2}\big(\phi_\tau^{(2)} + \phi_\tau^{(1)}\big) - \frac{1}{2}\big(\phi_\tau^{(2)} - \phi_\tau^{(1)}\big), \tag{23}$$

and

$$\phi_\tau^{(2)} = \frac{1}{2}\big(\phi_\tau^{(2)} + \phi_\tau^{(1)}\big) + \frac{1}{2}\big(\phi_\tau^{(2)} - \phi_\tau^{(1)}\big). \tag{24}$$

Substituting into Eq. (22) and working through some algebra, we have the generalized word shift form,

$$\delta\Phi = \sum_\tau \big(p_\tau^{(2)} - p_\tau^{(1)}\big)\left[\frac{1}{2}\big(\phi_\tau^{(2)} + \phi_\tau^{(1)}\big) - \Phi^{(\mathrm{ref})}\right] + \frac{1}{2}\big(p_\tau^{(2)} + p_\tau^{(1)}\big)\big(\phi_\tau^{(2)} - \phi_\tau^{(1)}\big). \tag{25}$$

The basic word shift [4] is a special case of the general form when the scores are text independent $\phi_\tau^{(1)} = \phi_\tau^{(2)} = \phi_\tau$,

$$\delta\Phi = \sum_\tau \big(p_\tau^{(2)} - p_\tau^{(1)}\big)\big(\phi_\tau - \Phi^{(\mathrm{ref})}\big). \tag{26}$$

### 6.3 Handling missing types and scores

At times, we may have a word present in one text only, and so either $p_\tau^{(1)} = 0$ and $p_\tau^{(2)} > 0$, or $p_\tau^{(2)} = 0$ and $p_\tau^{(1)} > 0$. If the scores $\phi_\tau^{(i)}$ are functions of the relative frequencies, then this can be problematic at times. Consideration needs to be given to the particular measure at hand to decide how to deal with the missing types. For some of the measures, like the generalized entropy, setting $\phi_\tau^{(i)} = 0$ does not cause any mathematical troubles. For the Shannon entropy, it may seem at first that we have cause for concern because $\phi_\tau^{(i)} = -\log p_\tau^{(i)}$, which is undefined if $p_\tau^{(i)} = 0$. However, in the Shannon entropy the surprisal is always multiplied by $p_\tau^{(i)}$, and so by the magic of limits and differential calculus, we can safely write

$$-p_\tau^{(i)} \log p_\tau^{(i)} = 0, \tag{27}$$

when $p_\tau^{(i)} = 0$. Practically, for programmatic implementation, it is enough to set $\phi_\tau^{(i)} = 0$. Similarly for the JSD, the form of $\phi_\tau^{(1)} = \log m_\tau / p_\tau^{(1)}$ may appear to be undefined when $p_\tau^{(1)} = 0$. If we step back to the overall word contribution $\delta\Phi_\tau$ shown in Eq. (6) though, then by the same limiting argument as the Shannon entropy, we can safely simply set $\phi_\tau^{(1)} = 0$. The same cannot be done for the KLD though, unfortunately. One part of the word contribution is the quantity $-p_\tau^{(2)} \log p_\tau^{(1)}$. Because the relative frequencies are different in this case (unlike the Shannon entropy and JSD), no amount of applying L'Hôpital's rule will give us a finite limit as $p_\tau^{(1)}$ approaches zero. All we can say is that the KLD is undefined when $p_\tau^{(1)} = 0$ and $p_\tau^{(2)} > 0$.

For dictionary scores, we may have both $p_\tau^{(1)} > 0$ and $p_\tau^{(2)} > 0$, but, without loss of generality, only $\phi_\tau^{(1)}$ is defined. This can happen in domain-adapted sentiment dictionaries. For example, if we were to build sentiment dictionaries for every year since the early 1900s, we would not be able to assign sentiment to the hashtag "#worldwar3" for texts from 1910 because neither social media nor World War I existed at the time. In a less extreme case, "trump" is certainly used enough in 2020 to be included in any contemporary frequency-based sentiment dictionary, but it may not have been used enough for inclusion in a dictionary for, say, the 1940s, even though the word certainly existed. There is ambiguity in how to handle these cases. We default to setting the missing score $\phi_\tau^{(2)}$ to the value of the defined score $\phi_\tau^{(1)}$, which we refer to as "borrowing" in the main text. In practice, this nullifies the score difference component, and places the contribution's emphasis on the basic word shift component. This may be reasonable in some cases and less so in others. It is questionable if we expect a shift in a word's sentiment to be well-defined and noticeable between two texts or contexts—as may be the case with "trump" between 1940 and 2020, for example. If that is the case, then it may be necessary to further expand the sentiment dictionaries through additional human or algorithmic annotation, or exert domain expertise to make some other defensible decision.

### 6.4 Word shift diagnostics

Generalized word shift graphs contain a significant amount of substantive and visual information. However, they are still a particular summary of the data, and so they can benefit from other summaries that contextualize the word contributions. We refer to these summaries as word shift diagnostics.

The first set of diagnostics are the bars present at the top of all word shift graphs, which show the cumulative contribution of each type of word shift and the total difference $\Sigma$. These summary bars quickly quantify the relative importance of each type of contribution. This is important because it provides context on what types of contributions are most significant, even if they do not appear among the top words presented in the word shift graph. For example, a word shift graph may report that the top fifty contributions come primarily from relatively negative words that appear more often ($-\uparrow$, dark blue bars to the left). However, the cumulative type contributions at the top may indicate that relatively negative words that appear more often are only a small portion of the *overall* sum of the other five types of word contributions. Without the summary cumulative bars at the top, it is difficult to quickly say what types of contributions, including those not shown in the top words, affect the word shift difference the most.

To further aid the interpretation of the cumulative effects, we include a second diagnostic plot that is in the bottom left corner of all of the case study plots in Sect. 4. We omit it from the presidential speeches word shift graphs for simplicity. This inset shows how each word cumulatively contributes to the total word shift difference as a function of rank. More specifically, we first rank all words by the absolute magnitude of their contribution $|\delta\Phi_\tau|$. We then cumulatively add these ranked absolute contributions to produce the cumulative contribution curve. If we normalize by the total sum of the absolute contributions,

$$\sum_{\tau \in \mathcal{T}} |\delta\Phi_\tau|, \tag{28}$$

we can interpret the plot as quantifying the percent of all variation explained up to a given word rank. The horizontal line in each plot marks the boundary between the words shown in the word shift graph and all those that are not. The intersection of the cumulative curve with the horizontal line indicates how much of the word shift difference is explained by the words presented in the figure. This is important for determining how much weight should be given to an interpretation that relies on the word shift graph. If a large portion of the variation is explained by the top words, then the word shift graph reliably describes most of the story of how two texts differ. However, if only a small portion of the variation is explained by the top words, as revealed by the cumulative contribution curve, then we can be less sure that the word shift graph is fully exposing the qualitative story. This may indicate that more words should be presented in the figure, or that additional analyses are needed for unpacking the texts, as we call for in our case study of 280 character tweets.

An alternative way that we may present the cumulative contributions, which we do not use in any of the word shift graphs here but that is available in the code we provide, is to plot the raw, rather than absolute, contributions as a function of rank. That is, plotting $\sum_\tau \delta\Phi_\tau$ as a function of rank, and normalizing by

$$|\delta\Phi| = \left|\sum_{\tau \in \mathcal{T}} \delta\Phi_\tau\right|. \tag{29}$$

This displays the trajectory of the word shift difference as we add additional words, which helps highlight effects of the long tail of contributions. This trajectory may be non-monotonic, unlike the absolute cumulative curve. Together with the total contribution bars at the top, the inset cumulative rank contribution plot gives us important summary information about how individual word contributions come together in total, and draws our attention to textual differences that may not be explained by the high ranking words that are visualized in the bar chart.

The final diagnostic plot, shown in the bottom right corner of most of the case study plots in Sect. 4, simply quantifies the relative size of the two corpora under study. The size is measured by counting the number of word tokens in each corpus. The text size diagnostic alerts us if one of our texts is much larger than the other. This can be especially problematic for any of the word shift measures that calculate their scores directly from the relative frequency distribution. If one text is much smaller than another, its word distribution will be less stable, and small differences may be improperly magnified by certain measures, like the Jensen–Shannon divergence [42].

## 6.5 Case studies

### *6.5.1 Presidential speeches*

We collected presidential speeches online from the University of Virginia's Miller Center (https://millercenter.org/the-presidency/presidential-speeches). The text of each speech is clearly organized by speaker and we parsed them to separate presidents from other entities (such as audiences or moderators). Unigrams were lowercased and the average sentiment was calculated over a president's entire set of speeches as a single text (not the average of average sentiments of each individual speech). Our dataset includes 71 speeches from Lyndon B. Johnson, consisting of 256,133 word tokens across 10,094 word types, and 39 speeches from George W. Bush, consisting of 107,913 tokens across 7804 types. For the labMT sentiment dictionary [4], we use a reference value of $\Phi^{(\mathrm{ref})} = 5$, which is the center of the dictionary's 1 to 9 sentiment scale. We also apply a stop window which excludes any labMT word whose sentiment falls between the scores 4 and 6. For the SocialSent historical lexicons [35], we use a reference value of $\Phi^{(\mathrm{ref})} = 0$, as all dictionaries were scaled to have a mean of zero when they were constructed.

### *6.5.2 Moby Dick*

The raw text of Moby Dick by Herman Melville is freely available on Project Gutenberg at http://www.gutenberg.org/files/2701/2701.txt. We process the raw text by removing the head matter and manually ending the text at the 'ETYMOLOGY' section. For the figures in this paper, we use a manually trimmed version of the raw text, with chapter headings removed (in contrast the larger emotional arc corpus [19], which relied on automated header and footer removal). We remove spaces and punctuation, and lowercased all tokens. There are 213,984 total tokens in Moby Dick across 16,858 word types, resulting in 106,992 tokens in each the first and second halves with 11,930 and 11,646 word types, respectively. For sentiment scores, we make the same choices as we did for the presidential speeches: we use the labMT sentiment dictionary [4], apply a stop window which excludes any labMT word whose sentiment falls between the scores 4 and 6, and use a reference value of 5. A reproducible analysis is available at https://github.com/andyreagan/shifterator-case-study-moby-dick (as mentioned above, the results herein rely on the 'raw' versions in the codebase).

### *6.5.3 US urban parks*

We collected tweets from Twitter's Decahose (10%) feed, stored in the Computational Story Lab's database at the University of Vermont. We restricted our sample to English language tweets with GPS coordinates posted from January 1st, 2012 to April 27th, 2015 (a period in which geolocation was widely used). Using boundaries from the US Census, we subsampled tweets within each of the 25 largest cities in the US by population. Within these cities, we found 297,494 posted within urban park boundaries using the Trust for Public Land's Park Serve database at https://parkserve.tpl.org/. To compare sentiment between in-park and out-of-park tweets, we paired each in-park tweet with the closest-in-time out-of-park tweet from another user within the same city (see ref. [57] for details).

Across the park tweets, there were 3,920,722 tokens across 451,627 word types. Across the out-of-park control tweets there were 3,861,357 tokens across 410,397 word types. For sentiment scores, we make the same choices as we did for the presidential speeches and Moby Dick: we use the labMT sentiment dictionary [4], apply a stop window which excludes any labMT word whose sentiment falls between the scores 4 and 6, and use a reference value of 5.

### *6.5.4 Information content of* 280 *character tweets*

We collected English-language tweets from Twitter's decahose (10%) feed, stored in the Computational Story Lab's database at the University of Vermont. Language detection came from the 'en' language label on each tweet provided by Twitter's API. This comprised 577,985,080 tweets over the 60-day period studied: 274,888,052 from the 30 days before 7 November 2017, and 303,097,028 from the 30 days afterwards. We restricted to considering changes in a consistent vocabulary of all 10,222 word types contained in the LabMT dictionary (i.e., without removal of any stop words) before and after the change. This resulted in a collection of 2,526,152,975 word tokens from the period before the change, and 2,555,503,284 from the period after the change. We use the average entropy of 140 character tweets as the reference value for the generalized word shift.

### *6.5.5 Regression analysis of urban labor diversity and the Great Recession*

Employment data for U.S. cities in 2007 comes from Occupational Employment Statistics (OES) data provided by the U.S. Bureau of Labor Statistics (BLS). Employment is reported using the Standard Occupation Classification (SOC) system that unifies occupational data across the U.S. Department of Labor. The SOC is a hierarchical classification system, and we use the most detailed (i.e., 6-digit) occupation codes in our analysis. However, occupation titles in Fig. 5 are simplified to conserve space; for example, the occupation category "Correctional Officers and Jailers" (occupation code: 33-3012) is simplified to "Correctional officers." For comparing high and low diversity cities in Fig. 5, we first rank U.S. cities based on the Shannon entropy of their employment distributions in 2007 (i.e., $H^{(c)}$) and consider the 15 most diverse cities to the 15 least diverse cities. For each one of these collections of 15 cities, we produce an aggregated employment distribution by taking the average employment share for each occupation across the cities in the collection of cities.

We analyze unemployment in U.S. cities during the Great Recession using Local Area Unemployment Statistics (LAUS) provided from the U.S. BLS. This data includes monthly statistics for each U.S. city. Since economic disruptions begin in different cities at different times and urban economies recover at different rates, we consider the month in the period between January 2008 and December 2012 with the most unemployment in a given city.

Table 2 displays a more complete analysis of urban labor statistics and unemployment during the Great Recession. All variables are centered and standardized prior to analysis so that each variable is unit-less; this makes it easier to compare the relative importance of each independent variables in predicting the dependent variable. It is very important to first control for the size of each city's labor force (i.e., $T^{(c)}$) before considering the effects of labor diversity on economic resilience. This is because cities with larger labor forces have greater potential for absolute unemployment. Models 1, 2, and 3 show the Pearson correlations between each individual independent variable and the dependent variable. Model 3 combines all independent variables and reveals that both $T^{(c)}$ and $H^{(c)}$ are significant predictors of maximum unemployment during the Great Recession, but occupation

**Table 2** Regressing urban labor statistics against $\log_{10}$ the maximimum unemployment in each U.S. city during the Great Recession. All variables are centered and standardized prior to analysis. Regression coefficient estimates from Model 4 are presented in Fig. 6

| Variable | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| $\log_{10}$ Total Employment ($T^{(c)}$) | 0.950*** | | | 1.214*** | 1.296*** |
| Employment Diversity ($H^{(c)}$) | | 0.802*** | | −0.220*** | −0.203*** |
| Occupation Diversity ($N^{(c)}$) | | | 0.925*** | −0.068 | −0.140 |
| $T^{(c)} \times H^{(c)}$ | | | | | 0.114 |
| $N^{(c)} \times H^{(c)}$ | | | | | −0.040 |
| $T^{(c)} \times N^{(c)}$ | | | | | −0.072* |
| $R^2$ | 0.903 | 0.643 | 0.856 | 0.914 | 0.915 |
| adj. $R^2$ | 0.903 | 0.642 | 0.856 | 0.913 | 0.914 |

$p_{\mathrm{val}} < 0.1^{*}$, $p_{\mathrm{val}} < 0.01^{**}$, $p_{\mathrm{val}} < 0.001^{***}$.

diversity (i.e., $N^{(c)}$) is not. Adding the measures for labor diversity in addition to labor force size yields an improvement in the overall predictive performance of the regression model from 90.3% variance explained to 91.4% thus accounting for an additional 14% of the unexplained variance when using labor force size alone. Finally, Model 5 includes the interaction terms between independent variables and again demonstrates the added predictive value of $H^{(c)}$ in addition to $T^{(c)}$. Interestingly, we also find large cities with large occupation diversity experienced lower unemployment during the Great Recession.

**Acknowledgements**
We thank Brandon Stewart and the other the attendees of the Ninth Annual Conference on New Directions in Analyzing Text as Data for their useful feedback on the presentation of word shift graphs. We also thank the editor and anonymous reviewers for helping us improve the clarity of our manuscript. Computations were performed on the Vermont Advanced Computing Core supported in part by NSF award No. OAC-1827314.

**Funding**
CMD and PSD are grateful for financial support from the Massachusetts Mutual Life Insurance Company and Google Open Source under the Open-Source Complex Ecosystems And Networks (OCEAN) project. Neither funding source had any role in the collection, analysis, or interpretation of the data.

**Abbreviations**
KLD, Kullback–Leibler divergence; JSD, Jensen–Shannon divergence; BLS, U.S. Bureau of Labor Statistics.

**Availability of data and materials**
Python code for producing word shift graphs is available at https://github.com/ryanjgallagher/shifterator under Apache License 2.0. It is also available as the Python package Shifterator through the Python Package Index (PyPi) at https://pypi.org/project/shifterator/. The version of the package specifically used to produce this manuscript is available at https://doi.org/10.5281/zenodo.4000536. Data for the presidential speeches case study was retrieved from the University of Virginia's Miller Center and is available online (https://millercenter.org/the-presidency/presidential-speeches). The raw text of *Moby Dick* was retrieved from Project Gutenberg and is freely available (http://www.gutenberg.org/files/2701/2701.txt). Data for the labor diversity case study comes from Occupational Employment Statistics data provided by the U.S. Bureau of Labor Statistics). The data for the 280 Character Tweets case studiy are from Twitter's Decahose (a 10% sample of tweets), and the data for the U.S. Urban Parks case study comes from Twitter's Spritzer (a 1% sample of tweets). According to Twitter's Terms of Service, raw tweet data cannot be shared. The Tweet IDs for both case studies are available upon request. U.S. urban park boundaries were identified according to data acquired through the Trust for Public Land's Park Serve (https://parkserve.tpl.org/).

**Competing interests**
The authors declare that they have no competing interests.

**Authors' contributions**
RJG and PSD derived the generalized word shift form and designed the presentation of word shift graphs. RJG implemented the core word shift graph code and wrote the initial draft of the manuscript. MRF, LM, AJS, and AJR all contributed case studies to the manuscript. CM and PSD advised the entire research process. All authors read, edited, and approved the final manuscript.

**Author details**
[1]Network Science Institute, Northeastern University, 02115 Boston, MA, USA. [2]Department of Informatics and Networked Systems, University of Pittsburgh, 15260 Pittsburgh, PA, USA. [3]Connection Science, Massachusetts Institute of

Technology, 02139 Cambridge, MA, USA. [4]Institute for Human-Centered Artificial Intelligence, Stanford University, 94305 Stanford, CA, USA. [5]School of Mathematical Sciences, The University of Adelaide, 5005 Adelaide, SA, Australia. [6]Computational Story Lab, Vermont Complex Systems Center, & Vermont Advanced Computing Core, The University of Vermont, 05401 Burlington, VT, USA. [7]Gund Institute for Environment & Rubenstein School of Environment and Natural Resources, The University of Vermont, 05401 Burlington, VT, USA. [8]Department of Ecology and Evolutionary Biology, University of Colorado at Boulder, 80309 Boulder, CO, USA. [9]MassMutual Data Science, 01002 Amherst, MA, USA. [10]MassMutual Center of Excellence for Complex Systems and Data Science & Department of Mathematics and Statistics, The University of Vermont, 05401 Burlington, VT, USA.

### Endnote

[a] The text of *Moby Dick* is freely available at http://www.gutenberg.org/files/2701/2701.txt.

## Publisher's Note

Received: 3 September 2020 Accepted: 6 January 2021 Published online: 19 January 2021

### References

1. Lazer D, Pentland A, Adamic L, Aral S, Barabási A-L, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M et al (2009) Computational social science. Science 323(5915):721–723
2. Salganik MJ (2019) Bit by bit: social research in the digital age. Princeton University Press, Princeton
3. Grimmer J, Stewart BM (2013) Text as data: the promise and pitfalls of automatic content analysis methods for political texts. Polit Anal 21(3):267–297
4. Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM (2011) Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. PLoS ONE 6(12):26752
5. Mitchell L, Frank MR, Harris KD, Dodds PS, Danforth CM (2013) The geography of happiness: connecting Twitter sentiment and expression, demographics, and objective characteristics of place. PLoS ONE 8(5):64417
6. 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. In: Proceedings of the 22nd international conference on the world wide web (WWW). ACM, New York, pp 307–318
7. Stella M, Ferrara E, De Domenico M (2018) Bots increase exposure to negative and inflammatory content in online social systems. Proc Natl Acad Sci USA 115(49):12435–12440
8. Barron AT, Huang J, Spang RL, DeDeo S (2018) Individuals, institutions, and innovation in the debates of the French Revolution. Proc Natl Acad Sci USA 115(18):4607–4612
9. Petersen AM, Tenenbaum J, Havlin S, Stanley HE (2012) Statistical laws governing fluctuations in word use from word birth to word death. Sci Rep 2:313
10. Pechenick EA, Danforth CM, Dodds PS (2017) Is language evolution grinding to a halt? The scaling of lexical turbulence in English fiction suggests it is not. J Comput Sci 21:24–37
11. Sims M, Bamman D (2020) Measuring information propagation in literary social networks. arXiv:2004.13980
12. Kolchinsky A, Dhande N, Park K, Ahn Y-Y (2017) The minor fall, the major lift: inferring emotional valence of musical chords through lyrics. R Soc Open Sci 4(11):170952
13. Funk K, Mullen LA (2018) The spine of American law: digital text analysis and us legal practice. Am Hist Rev 123(1):132–164
14. Voigt R, Camp NP, Prabhakaran V, Hamilton WL, Hetey RC, Griffiths CM, Jurgens D, Jurafsky D, Eberhardt JL (2017) Language from police body camera footage shows racial disparities in officer respect. Proc Natl Acad Sci USA 114(25):6521–6526
15. Alajajian SE, Williams JR, Reagan AJ, Alajajian SC, Frank MR, Mitchell L, Lahne J, Danforth CM, Dodds PS (2017) The lexicocalorimeter: gauging public health through caloric input and output on social media. PLoS ONE 12(2):0168893
16. Pechenick EA, Danforth CM, Dodds PS (2015) Characterizing the Google Books corpus: strong limits to inferences of socio-cultural and linguistic evolution. PLoS ONE 10(10):0137041
17. Gallagher RJ, Reagan AJ, Danforth CM, Dodds PS (2018) Divergent discourse between protests and counter-protests: #BlackLivesMatter and #AllLivesMatter. PLoS ONE 13(4):0195644
18. Dodds PS, Danforth CM (2009) Measuring the happiness of large-scale written expression: songs, blogs, and presidents. J Happ Stud 11(4):441–456. https://doi.org/10.1007/s10902-009-9150-9
19. Reagan AJ, Mitchell L, Kiley D, Danforth CM, Dodds PS (2016) The emotional arcs of stories are dominated by six basic shapes. EPJ Data Sci 5(1):31
20. Baylis P, Obradovich N, Kryvasheyeu Y, Chen H, Coviello L, Moro E, Cebrian M, Fowler JH (2018) Weather impacts expressed sentiment. PLoS ONE 13(4):0195750
21. Nelson LK, Burk D, Knudsen M, McCall L (2021) The future of coding: a comparison of hand-coding and three types of computer-assisted text analysis methods. Sociol Methods Res 50(1):202–237
22. Muddiman A, McGregor SC, Stroud NJ (2019) (Re) claiming our expertise: parsing large text corpora with manually validated and organic dictionaries. Polit Commun 36(2):214–226
23. Brady WJ, Wills JA, Jost JT, Tucker JA, Van Bavel JJ (2017) Emotion shapes the diffusion of moralized content in social networks. Proc Natl Acad Sci USA 114(28):7313–7318
24. Sood S, Antin J, Churchill E (2012) Profanity use in online communities. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 1481–1490
25. Schmidt A, Wiegand M (2017) A survey on hate speech detection using natural language processing. In: Proceedings of the fifth international workshop on natural language processing for social media, pp 1–10
26. Loughran T, McDonald B (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. J Finance 66(1):35–65

27. Barberá P, Boydstun AE, Linn S, McMahon R, Nagler J (2021) Automated text classification of news articles: a practical guide. Polit Anal 29(1):19–42
28. Zipf GK (1949) Human behavior and the principle of least effort. Addison-Wesley, Reading
29. Simon HA (1955) On a class of skew distribution functions. Biometrika 42(3/4):425–440
30. Pury CL (2011) Automation can lead to confounds in text analysis: Back, Küfner, and Egloff (2010) and the not-so-angry Americans. Psychol Sci 22(6):835
31. Schmidt BM (2012) Words alone: dismantling topic models in the humanities. J Dig Humanit 2(1):49–65
32. Munro R (2010) Crowdsourced translation for emergency response in Haiti: the global collaboration of local knowledge. In: Proceedings of the AMTA workshop on collaborative crowdsourcing for translation, pp 1–4
33. Schwaiger JM, Lang M, Ritter C, Johannsen F (2016) Assessing the accuracy of sentiment analysis of social media posts at small and medium-sized enterprises in Southern Germany
34. Bucholtz M, Bermudez N, Fung V, Edwards L, Vargas R (2007) Hella nor cal or totally so cal? The perceptual dialectology of California. J Eng Linguist 35(4):325–352
35. Hamilton WL, Clark K, Leskovec J, Jurafsky D (2016) Inducing domain-specific sentiment lexicons from unlabeled corpora. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP), vol 2016, p 595
36. Baucom E, Sanjari A, Liu X, Chen M (2013) Mirroring the real world in social media: Twitter, geolocation, and sentiment analysis. In: Proceedings of the 2013 international workshop on mining unstructured big data using natural language processing. ACM, New York, pp 61–68
37. Reagan AJ, Danforth CM, Tivnan B, Williams JR, Dodds PS (2017) Sentiment analysis methods for understanding large-scale texts: a case for using continuum-scored words and word shift graphs. EPJ Data Sci 6(1):28
38. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379–423
39. Havrda J, Charvát F (1967) Quantification method of classification processes: concept of structural *a*-entropy. Kybernetika 3(1):30–35
40. Jost L (2006) Entropy and diversity. Oikos 113(2):363–375
41. Hill MO (1973) Diversity and evenness: a unifying notation and its consequences. Ecology 54(2):427–432
42. Altmann EG, Dias L, Gerlach M (2017) Generalized entropies and the similarity of texts. J Stat Mech Theory Exp 2017(1):014002
43. Steiner M, Magin M, Stark B, Geiß S (2020) Seek and you shall find? A content analysis on the diversity of five search engines' results on political queries. Inf Commun Soc. https://doi.org/10.1080/1369118X.2020.1776367
44. Mohammad S (2018) Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In: Proceedings of the 56th annual meeting of the association for computational linguistics (ACL), vol 1, pp 174–184
45. Mohammad SM (2018) Word affect intensities. In: Proceedings of theition of the language resources and evaluation conference (LREC-2018)
46. Dodds PS, Clark EM, Desu S, Frank MR, Reagan AJ, Williams JR, Mitchell L, Harris KD, Kloumann IM, Bagrow JP, Megerdoomian K, McMahon MT, Tivnan BF, Danforth CM (2015) Human language reveals a universal positivity bias. Proc Natl Acad Sci 112(8):2389–2394
47. Denny MJ, Spirling A (2018) Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. Polit Anal 26(2):168–189
48. Schwartz AJ, Dodds PS, O'Neil-Dunne JP, Danforth CM, Ricketts TH (2019) Visitors to urban greenspace have higher sentiment and lower negativity on Twitter. People Nat 1(4):476–485
49. Ihara I (2017) Our discovery of cramming. Twitter Developer Blog. https://blog.twitter.com/engineering/en_us/topics/insights/2017/Our-Discovery-of-Cramming.html
50. Neubig G, Duh K (2013) How much is said in a tweet? A multilingual, information-theoretic perspective. In: 2013 AAAI spring symposium series
51. Rosen A (2017) Tweeting made easier. Twitter Developer Blog. https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html
52. Mitchell L, Dent J, Ross JV (2018) Mo' characters mo' problems: online social media platform constraints and modes of communication. AoIR Selected Papers of Internet Research
53. Elsby MW, Hobijn B, Sahin A (2010) The labor market in the Great Recession. Technical report, National Bureau of Economic Research
54. Oliver TH, Heard MS, Isaac NJ, Roy DB, Procter D, Eigenbrod F, Freckleton R, Hector A, Orme CDL, Petchey OL et al (2015) Biodiversity and resilience of ecosystem functions. Trends Ecol Evol 30(11):673–684
55. Frank MR, Sun L, Cebrian M, Youn H, Rahwan I (2018) Small cities face greater impact from automation. J R Soc Interface 15(139):20170946
56. Dodds PS, Minot JR, Arnold MV, Alshaabi T, Adams JL, Dewhurst DR, Gray TJ, Frank MR, Reagan AJ, Danforth CM (2020) Allotaxonometry and rank-turbulence divergence: a universal instrument for comparing complex systems. arXiv:2002.09770
57. Schwartz AJ, Dodds PS, O'Neil-Dunne JPM, Ricketts TH, Danforth CM (2020) Gauging the happiness benefit of US urban parks through Twitter. arXiv:2006.10658