Journal of Economic Growth (2022) 27:415–452
https://doi.org/10.1007/s10887-022-09204-6

How does scientific progress affect cultural changes? A digital text analysis

Michela Giorcelli[1] · **Nicola Lacetera**[2] · **Astrid Marinoni**[3]

Accepted: 2 March 2022 / Published online: 23 April 2022

Abstract
We study the effects of scientific changes on broader cultural discourse, two phenomena that the economics literature identifies as key drivers of long-term growth, focusing on a unique episode in the history of science: the elaboration of the theory of evolution by Charles Darwin. We measure cultural discourse through the digitized text analysis of a corpus of hundreds of thousands of books as well as of Congressional and Parliamentary records for the US and the UK. We find that some concepts in Darwin's theory, such as Evolution, Survival, Natural Selection and Competition, significantly increased their presence in the public discourse immediately after the publication of *On the Origin of Species*. Moreover, several words that embedded the key concepts of the theory of evolution experienced semantic and sentiment changes—further channels through which Darwin's theory influenced the broader discourse. Our findings represent the first large-sample, systematic quantitative evidence of the relation between two key determinants of long-term economic growth, and suggest that natural language processing offers promising tools to explore this relation.

Keywords Culture · Science · Natural language processing · History · Growth

JEL Classification C55 · N00 · O39 · O49 · Z10 · Z13

✉ Nicola Lacetera
nicola.lacetera@utoronto.ca

Michela Giorcelli
mgiorcelli@econ.ucla.edu

Astrid Marinoni
astrid.marinoni@scheller.gatech.edu

[1] UCLA Department of Economics, and NBER, University of California, 9262 Bunche Hall, 315 Portola Plaza, Los Angeles, CA 90095, USA

[2] University of Toronto, and NBER, 3359 Mississauga Road North, Mississauga, ON, Canada

[3] Scheller College of Business - Georgia Institute of Technology, 800 W Peachtree St NW, Atlanta, GA, USA

1 Introduction

"Economic change in all periods depends, more than most economists think, on what people believe." (Joel Mokyr, The Enlightened Economy)
"Every historical act can only be performed by the 'collective man', and this presupposes the attainment of a 'cultural-social' unity […], on the basis of an equal and common conception of the world." (Antonio Gramsci, The Prison Notebooks)

As the scholarly quest for the determinants of economic growth shifted attention away from factors such as labor, land and capital, a large literature identified scientific and technological progress as a key driver of development and prosperity (Bush, 1945, Jones 2002, Pakes & Sokoloff, 1996, Romer, 1990, Stephan, 2012). In the last few decades, scholars also pointed to the role of culture, i.e., the set shared beliefs, values, goals and traditions that a population holds and transmits over time, as a further determinant of the institutional choices and economic trajectories of a community (Alesina & Giuliano, 2015, Galor, 2011, Guiso et al., 2006, Landes, 1999, McCloskey, 2016, Mokyr, 2016, Spolaore 2014, 2020, Williamson, 2000)

We know little, however, about the *relationship* between science and culture. If they do not only develop independently but also interact with each other, this relationship may represent a further variable of interest to understand economic change. Mokyr (2013, 2016), for example, advanced the idea that certain scientists introduce new sets of beliefs in a population with their discoveries.[1] The impact of these individuals, therefore, affects not only the production and diffusion of scientific knowledge, but also changes how people, more broadly, interpret the world around them.[2] Mokyr calls these scientists "cultural entrepreneurs".

In this paper, we propose an approach to test empirically the impact of scientific progress on the broader culture, and we apply our methodology to one of the major advancements in the history of science: Charles Darwin's theory of evolution by natural selection. Assessing how scientific progress affects cultural change presents several empirical challenges. First, one would need a long time horizon to analyze the interplay between public discourse and scientific progress. Second, unobserved factors and events (especially over extended periods) make inferring causal links difficult. A further complication is how to define and measure, in the first place, culture and cultural change.

Most of the existing literature in the economics of culture relies on survey-based measure of specific attitudes, such as trust, cooperation, or "civicness", or on activities whose intensity plausibly correlates with some of those attitudes.[3] To our knowledge, there have not been attempts to define measures of culture related to new scientific ideas and discoveries. Moreover, most existing measures concern recent times and represent beliefs in a given moment.[4] To conceptualize and measure culture and cultural change over a long (historical) period, and to identify the inclusion of certain scientific ideas into the broader

[1] Mokyr (2013), moreover, distinguishes between "macro" and "micro" scientific discoveries. He argues that only the former creates a discontinuous change. The latter are important to guarantee continuous improvements, but do not cause any scientific or cultural breakthrough.
[2] Similarly, Schiller (2017) advances the idea that certain individuals may introduce "narratives" that spread in a society and affect broader beliefs and consequent actions.
[3] See for example Guiso et al. (2004).
[4] Giuliano and Nunn (2021) have recently advanced an agenda to measure cultural persistence and change.

public discourse, we adopt a different approach. We rely on concepts, sources and tools from such fields of research as the humanities, sociology and ethnography. The underlying claim of these approaches is that language embodies values and beliefs, and is a major channel of communication and transmission of them over time (Hamilton et al 2016; Kirby et al., 2007; Lévi-Strauss 1963; Nguyen et al., 2020; Whorf 1956). Changes in use of certain words and phrases, as well as in the meaning of a word, indicate changes in underlying beliefs and views of the world in ways that can be transmitted and further change (in a measurable way).

We study the evolution of certain phrases and expressions by performing digital text analysis on a large text corpus between 1820 and 1899. Written text, of course, represents only a part language-based communications, together with oral exchanges (Michalopoulos & Xue, 2021). There are, moreover, relevant forms of non-verbal behaviors and communications as well. In addition to being, in general, easier to record and measure the written word, the growth of the printing industry and the increase in literacy rates especially since the 19th Century, and the role not only of academic and other non-fiction texts, but also of the fictional literature especially through the diffusion of the novel (Lyons, 2003), makes the written language a major repository (and means of transmission) of broader values and beliefs. In fact, a relevant claim in the digital humanities and cultural linguistics literature is that digital text analysis or "distant reading" allows for the consideration of the "great unread", i.e., the large quantity of texts that normally scholars do not study, but that, as a whole, represent the broader social and cultural climate and discourse at a given time (Cohen, 1999).

Although there is a general perception that the theory of evolution had broader influence, we know less about what concepts were particularly influential, how their influence evolved and entered the public discourse, and how long it took these ideas to diffuse beyond a narrow scientific community. The information that we retrieve from various text sources makes progress in addressing these questions.

The publication of *On the Origin of Species*, in 1859, made Darwin's theory known to a vast public; moreover, the timing of the publication was largely unplanned. We rely on this event as our main source of natural variation. The main corpus on which we perform our text analysis is Google Books, a digitized collection of about eight million volumes. We define the publication year of *On the Origin of Species* as our reference date and concentrate the analysis on the four decades before and after it (1820–1899). We consider words and expressions that, according to many accounts, embody the key concepts of Darwin's theory (Desmond & Moore, 1994; Mayr, 1982): Evolution, Survival, Competition, (Natural) Selection, and Adaptation. We compare the evolution of the frequency of use of these Darwinian words with a large number of words not related directly to Darwin's theory but extensively used in *On the Origin of Species*. We then complement the frequency analysis on the Google Books corpus with evidence from UK Parliamentary Transcripts and US Congressional Records. With these additional corpora, we explore how certain concepts diffused not only in the cultural discourse, but also in the political arena, thus potentially shaping the policy debate. In addition to frequency of use, we assess semantic changes and the evolution of attitudes toward Darwinian concepts as additional measures of cultural change, applying word-embedding techniques.[5]

[5] We limit the semantic and sentiment analysis to the Google Book corpus because of the demanding sample size requirements of the underlying methodologies.

We show, first, that some key concepts in Darwin's theory increased their diffusion in the broader cultural discourse in the years immediately following the publication of *On the Origin of Species*: Natural Selection, Evolution, Survival and Competition. The patterns of diffusion of these words and expressions were similar in the non-fiction and fiction literature; this indicates that the underlying concepts had a broad impact on culture as well as on the social imaginary as represented, for example, by short stories and novels. Other concepts such as Selection and Adaptation did not experience a change in the rate of diffusion. We also document that some of the key Darwinian ideas entered the policy debate after 1859 but with some delay with respect to the entry of these concepts in the broader public discourse. The effects of *On the Origin of the Species*, moreover, were not specific to the English-speaking world; the Darwinian concepts diffused in non-English speaking countries right after the translation of the book in the corresponding language. Moreover, the translation occurred earlier in countries that industrialized earlier, such as Germany and France, than in "late comers" such as Italy and Spain.

The second set of results concerns changes in the semantics of these words as well as in the types of reactions, or sentiments, that they generated over time. Of interest is, for example, the increase in semantic association between certain words, such as Competition and Life, as well as between Life and Adaptation. The term Evolution, which came mostly from chemistry and physics in the first half of the 1800s, later in the century related more to concepts from biology as well as social and human subjects, indicating a broader reach of this idea in society. We also document an increased similarity between Evolution and words related to the traditional view of the Christian doctrine about the origins of the world, such as Creation and Genesis; this suggests a process of "secularization" of these ideas. Furthermore, Selection became more similar in meaning to other "Darwinian" words, such as Survival, Variation, Fittest and Heredity. Sentiment analysis shows a more positive attitude toward certain Darwinian concepts after the publication of *On the Origin of Species*, in particular Evolution, and a positive attitude toward Darwin himself.

Finally, we show that the word "Darwin" diffused more literature than the names of other major scholars in the same area (Lamarck, Chambers and Wallace), and that the semantic association of the focal concepts that we consider was higher with the name "Darwin" than with the other names. This suggests that these ideas were particularly associated, in the public discourse, with Darwin's work and not just generically with the progress in the biological sciences of the time or ideas that were "in the air".

The relationship between scientific discoveries and the public discourse may also contribute to understanding deeper social and political processes, such as the extent to which, to cite Alexander Hamilton's reflections in the *Federalist Papers*, a society is based on a "culture of reason and evidence". If a culture that values scientific inquiry is more likely to promote economic development, and scientific breakthroughs contribute to the evolution of culture in this direction, then studying this relationship acquires additional value. We see our approach as a fruitful one to investigate also the impact of other scientific breakthroughs in history.

In Sect. 2, we provide a brief account of Darwin's elaboration of the theory of evolution by natural selection. We also explain why the publication of *On the Origin of Species* provides natural variation that allows studying the effect of Darwin's theory on the broader public discourse. In Sect. 3, we describe the text-based data that we use and the techniques and empirical strategies that we adopt to extract information about cultural change. Section 4 reports the findings. In Sect. 5, we provide a discussion and propose directions for future research.

2 Historical background and identification

"It is doubtful if any single book, except the 'Principia', ever worked so great and so rapid a revolution in science, or made so deep an impression on the general mind."
Obituary for Charles Darwin, Proceedings of the Royal Society of London, 1888.

2.1 The development of Darwin's theory of evolution

Charles Darwin's interest in the evolution of living organisms largely developed during his voyage on the HMS Beagle, a ship of the Royal Navy, from 1831 to 1836. Over those five years, Darwin collected fossils from the places that he visited and observed their geographical distribution. Although his early conjectures built on previous theories (such as Lamarck's and Chambers') and considered the possibility of the transformation of one species into another (transmutation), he then developed his own theory of evolution based on the natural selection of the most adaptive (innate) characteristics of a species. Small, gradual variations within a species would emerge randomly, and would lead to branching of new species. Competition for resources and adaptive capacities would determine whether and where a particular species would be more likely to thrive. The developments in genetic research in 20th century provided corroboration and foundations to Darwin's theory (Desmond & Moore, 1994; Mayr, 1982).[6]

In addition to being one of the greatest scientific breakthroughs in history, there is a perception that Darwin's theory of evolution had a wider cultural reach (Desmond & Moore, 1994; Fuller, 2017; Mayr 1982, 2001). Research in literary criticism analyzed how the production of certain poets and novelists began to reflect the competition and "struggle" for resources, the common origins of species (including humans), and a new conception of the role of nature and God in the creation.[7] Mokyr (2013, 2016) includes Darwin among a small set of "cultural entrepreneurs", i.e., scientists whose discoveries affected deeply held and broadly shared popular beliefs. These accounts, however, focus on a narrow set of literary contributions or debates mostly restricted to scientific, political and economic elites, or a few highly successful literary works; this makes it hard to advance inferences about the broader cultural impact of this scientific advance, and about the cultural climate that preceded that breakthrough. Our approach to answering these questions, based on large text corpora, allows going beyond the analysis of a small set of texts and authors as a way to extrapolate general cultural views and trajectories.

2.2 The publication of *On the Origin of Species* as a source of natural variation

Some features of how Darwin made his work public enable us to identify the impact of his work on the broader cultural discourse. Although Darwin developed his theory over a long

[6] See in particular Desmond and Moore (1994) for details on the personal and intellectual biography of Darwin.

[7] Similarly, studies of the literary production prior to the publication of *On the Origin of Species* point out how some of Darwin's ideas connected to images already developed by these writers. A frequently cited example is the work of Alfred Tennyson, and in particular his poem *In Memoriam*, published in 1850. Scholars also investigated the connections between broader worldviews, such as Enlightenment and Romanticism, on Darwin's ideas (Cartwright and Baker 2005; Chapple 1986; Gianquitto and Fisher 2014; Lansley 2016; Otis 2009; Richards 2013; Scholnick 2015).

period, there is a precise time at which Darwin's theory reached the broader public, and this is 1859, the year of publication of *On the Origin of Species*.[8] This publication date was largely unplanned. Darwin proceeded slowly initially and had to deal with sickness and deaths in his family that further delayed him. However, eventually he "rushed" in order not to lose priority over Alfred R. Wallace, who was researching on the same topics and had sent Darwin some of his writings that developed similar concepts and reached similar conclusions about natural selection.

The book and Darwin's theory received almost immediate attention and diffusion, thanks to presentations at scientific meetings such as the Linnaean Society (of a joint paper with Wallace in 1858) and the British Association for the Advancement of Science (in 1860), as well as reviews in the popular press (see for example Gray, 1860; Huxley, 1859).

The unplanned publication date of Darwin's theory provides the main source of variation for our empirical study. The rapid diffusion of the theory gives us an opportunity to observe the effect on the diffusion of the main concepts, and to establish which ones were especially novel and had an independent impact on the broader public discourse.

To be sure, *On the Origin of Species* was not the first treatment of evolution. Darwin's theory was novel in several ways and more coherent than previous ones. However, earlier in the 19th Century some related ideas were already elaborated and discussed; examples include the work of Lamarck, the anonymous *Vestiges of the Natural History of Creation* (later attributed to the Scottish journalist and publisher Robert Chambers), and of course the work of Alfred R. Wallace. Herbert Spencer, moreover, published his *Principles of Biology*, which apply some Darwinian concepts also to society and ethics and not only to the natural sphere, in 1864. Our empirical strategy, however, allows assessing whether the publication of Darwin's book represented a discontinuous change in the cultural discourse. In the analyses reported below we also attempt to address the issue of whether the theory of evolution as Darwin presented it was already "in the air" with a variety of empirical exercises.

3 Data and methods

> *"The limits of my language mean the limits of my world."* Ludwig Wittgenstein, Tractatus Logico-Philosophicus (1922).

To examine the diffusion and the evolution of the meaning and interpretation of scientific concepts over time, we exploit the increasing availability of digitized text corpora, as well as the tools of natural language analysis. We rely on recent work at the intersection of the humanities, linguistics, cultural studies and computer science, which uses the frequency of use of words in large text corpora and their semantic evolution as measures of changes in the public discourse and shared beliefs. Relying on natural language processing techniques, this line of research has explored, for example, the evolution of cultural trends

[8] The year 1859 saw also the publication of other important works, John Stuart Mill's *On Liberty*, Tennyson's *Idylls of the King*, Eliot's *Adam Bede* and Dickens' *A Tale of Two Cities*. These publications make it harder to identify a connection between the publication of *The Origins of Species* and changes in the public discourse. However, in our study, we focus on specific concepts that are central in Darwin's work but not in the other works mentioned above; we also consider the presence of those concepts in the public discourse *before* 1859. Below, we also assess the presence of any abnormal trend in the number of words in our dataset, and in the introduction of new words, around 1859.

as expressed by both the frequency of use of certain words and phrases and the change of their meaning, the evolution of literary styles, and scholarly influence on various areas of research.[9]

These approaches focus on a particular source of expression and communication of cultural beliefs, i.e., formal language especially through the written word. This excludes more informal, but not less important, means of development and transmissions of values and worldviews. The increasing availability of text in digital forms for longer time periods, and the comparability of written text over time because of its "standardized" nature, provide, despite the limitations, a fruitful direction to learn, measure and compare ideas, beliefs and values from the past. In fact, also research in economics and economic history is increasingly using text as a source of data. Kelly et al. (2021), for example, apply text analysis techniques to patent data to build new measures of scientific and technological breakthrough. Michalopoulos and Xue (2021) rely on transcriptions of oral communications, such as folktales, and relate them to specific cultural traits such as trust, risk-aversion, and gender norms across countries and ethnicities.[10]

By operationalizing concepts with certain words and phrases, our specific objective is to document which ideas emerged as novel in society following our scientific breakthrough of interest, and whether they influenced different cultural spheres over time. The first step in this investigation is to compute relative frequencies of some key words that embody the main concepts in Darwin's theory of evolution, and that Darwin used extensively in his own work. These frequencies represent a basic measure of the adoption of certain ideas in the broader cultural and social discourse. Our investigation then focuses on word embeddings for the analysis of semantic and sentiment change (Aiden & Michel, 2014; Manovich, 2009; Michel et al., 2011; Roth, 2014).

3.1 Word frequencies

We first rely on Google N-Grams[11] (Lin et al., 2012) to assess how frequencies of words changed over time in fiction and non-fiction literature. The Google N-Grams data is a result of the Google Book project to build a vast collection of digitized books in partnership with major libraries.[12] First released in 2010, the data consist of a set of corpora of roughly eight million books, an estimated 6% of all books ever published (Lin et al., 2012). The texts cover roughly a 500-year span and there is a continuous update. The database includes different languages (besides English: Italian, French, German, Spanish, Russian, Hebrew, and Chinese). The English corpus alone has half a trillion words in it. For the period that we consider (i.e., 1820–1899), there are about 380,000 books containing more than 45 billion words in total.[13] The data include both fiction and non-fiction books, but not periodicals,

[9] Aiden and Michel (2014), Gerow et al. (2018), Hamilton et al., (2016a, 2016b), Heuser and Le-Khac (2011), Heuser (2016), Kozlowski et al. (2019), Manovich (2009), Michel et al. (2011), Moretti (2013), Thompson et al. (2020), Wilkens (2015).

[10] For various applications in political economy, the study of media, innovation, marketing and finance, see also Balsmeier et al. (2018), Bandiera et al. (2017), Catalini et al. (2015), D'Amico and Tabellini (2017), Enke (2020), Iaria et al. (2018), Jelveh et al. (2014), Kearney and Liu (2014), Kozlowski et al. (2019), Yin et al. (2021). Gentzkow et al. (2018) provide a survey of the use of text as data in economics.

[11] Available at: http://books.google.com/ngrams.

[12] http://books.google.com/googlebooks/library/partners.html.

[13] Over this period, there are about 1.26 million unique words on average per year.

and is aggregated depending on the number of terms considered; for instance, the 1-Ngram dataset includes single words and their frequency in a given corpus, and *n*-grams are combinations of *n* words and their frequency. We compute frequencies from 1-Ngrams and 2-Ngrams data for each year and express them in per-million-words terms.

The ability to separate fiction and non-fiction literature is relevant to us for two reasons. First, one critique to the N-Grams (and Google Books) corpus is that it may over-represent scientific texts (Pechenick et al., 2015). In our study, increases in the frequency of words related to Darwin's theory may just reflect a disproportionate increase over time of the corpus of scientific books (included in the non-fiction category). Second, separating fiction and non-fiction literature enables the analysis of different types of relationships between Darwinian science and broader culture. The use of Darwin's concepts in the non-fiction literature may better represent higher-educated or more erudite conversations. Conversely, given the diffusion of the novel, including in low-middle classes, and the relatively high literacy rates especially in England and the United States in the 19th century (Lyons, 2003), fictional literature may better measure the social imaginary (Armstrong, 1987; Winans, 1975).

Whilst the Google Books data allow us to measure Darwin's influence on the broader cultural discourse, we also aim to assess whether his ideas diffused in the political discourse – and thus, potentially, in the policy process. We rely on the digitized collections of the UK Parliamentary Debates (Hansard) and the U.S. Congressional records (ProQuest's Congressional Record Permanent Digital Collection).[14] The former includes reports of all discussions occurring in the House of Commons and House of Lords[15]; the latter focuses on debates in the House and Senate.[16]

3.2 Word meaning and embeddings

The analysis of word frequencies is informative, but does not provide insights about how a given word was used and its perception in society. The semantic changes and the evolution of attitudes toward a concept may be a more appropriate measure of cultural change if one interprets the meaning of a word as the association of that word with other concept and ideas, and the attitudes toward a concept as whether that concept had a positive or negative reception.

Natural language processing employs word-embedding techniques to determine the meaning of, and sentiment toward words from large text corpora, and their evolution over time. The main idea of word embedding is that we can evaluate semantic associations between words by analyzing co-occurrence patterns in a text. Two words of similar meaning are unlikely to appear, say, in the same sentence, but they are likely to be surrounded by similar words. For example, we would not expect that, within five words before and after the word "queen", we read the word "monarch"; however, there is plausibly high overlap

[14] See, among others, Fetter (1975) and Gentzkow and Shapiro (2010) for studies that relied on text from Congressional and Parliamentary records.

[15] Parliamentary Debates (Series 1 to 4 – 1903 to 1908).

[16] These include Congressional Record (1873–1997), the Congressional Globe (1833–1873), the Register of Debates in Congress (1824–1837), and the Annals of Congress (1789–1824). It is worth noticing that before 1873 each House was only required to keep an internal journal of its proceedings. Only from 1874 onwards were external reporters allowed to witness debates and granted full permission to report them (McPherson 1942).

between the words that appear immediately before and after "queen", and those that appear before and after "monarch".

The outcome of word-embedding algorithms is a set of vectors that include information about co-occurring patterns among words. Consider for example a text corpus with V words w ($w = 1, 2, \ldots, V$). For each word, one can specify a subset of "context words". i.e., terms that appear within a window of m words before and after w. The objective is to represent each word w as a Nx1 vector, with $N < V$ determined by the researcher, where each entry is a measure of how frequent the occurrence of w with each of the context words is. We rely on the Word2Vec approach (SkipGram with negative sampling; Mikolov et al., 2013), a technique that studies of semantic change have used extensively (e.g., Bolukbasi et al., 2016; Caliskan et al., 2017; Garg et al., 2017). The Word2Vec model is based on a neural-network structure that we represented, in simplified form, in Fig. 1A. The starting point is the definition of Vx1 one-hot vectors for each focal word w, i.e. vectors of all 0's except one value of and entry of 1 in correspondence of the word of interest. Two matrices, called the embedding and context matrices, are initially filled with random weights that the training process updates. For each word w, the algorithm multiplies a one-hot vector, or input layer, by the embedding VxN matrix, to obtain a Nx1 vector, called the hidden layer. This vector simply "copies" the input layer into the embedding matrix that corresponds to the word w. In turn, multiplying the hidden layer by the NxV context matrix produces the $Vx1$ output layer. The V entries (or scores) in the output layers go through a soft-max activation function, which maps the scores to a probability distribution. The probability vectors have values that range from 0 to 1 and sum up to 1.[17]

These vectors can now be compared to the "target" one-hot-encoding vector of a given context word c to obtain a vector of errors by subtracting the probability vector from the "target" vector. Using this information, a backpropagation mechanism (Rumelhart et al., 1986) updates the weights in the embedded and context matrix. The training process proceeds by considering all combinations of words w and context words c.

The final output consists of a VxN "embedding matrix". Each row in the matrix is the vector representation of each of the V words w, where each entry is a coordinate in an N-dimensional space and carries information about the context. The embedded vectors satisfy some "linearity" features in the relationship between, for example, the singular and plural form, or feminine and masculine version, of a word. Using a frequent example in the literature, we expect that, when the word vectors corresponding to king, kings, queen, queens, man and woman, the following holds: *(king–kings)* ≈ *(queen–queens)* and *(king–man)* ≈ *(queen–woman).*

The closer two word vectors are in this N-dimensional space, the stronger the semantic association between the two words. The main metric of the proximity between vectors is the cosine between them (Dubossarsky et al., 2015; Gulordava & Baroni, 2011; Jatowt & Duh, 2014; Kim et al., 2014; Kulkarni et al., 2015). Call γ the angle between two N-dimensional vectors $u = (u_1, \ldots u_N)$ and $v = (v_1, \ldots v_N)$. Then, $u'v = \sqrt{\sum_{i=1}^{N} u_i^2} * \sqrt{\sum_{i=1}^{N} v_i^2} * \cos(\gamma) = \|u\|\|v\|\cos(\gamma)$, or: $\cos(\gamma) = \frac{u'v}{\|u\|\|v\|} \in [-1, 1]$. The more similar the two vectors, the closer to one the cosine.

[17] The softmax function maps scores into probability distributions as follows: $p(c|w;\theta) = \frac{e^{v_c v_w}}{\sum_{c' \in C} e^{v' c v_w}}$, where v_c and v_w are vector representations of context word c and focal word w respectively, and C is the set of all possible contexts. The estimation procedure thus consists of maximizing the probability that a given context word occurs within a given window around each focal word of interest.

We investigate whether the words that defined the main Darwinian concepts shared context words with different terms before and after the publication of *On the Origin of Species*. We rely on previously trained Word2Vec embeddings resulting from the N-grams distributed by Google Books (). Figures are available for every decade between 1800 and 1990 and data are designed to enable comparisons across decades. The models use a context window of four context words and parameters as suggested by Levy et al. (2015) to measure semantic changes in cultural shifts.

The measure of semantic similarity, and more generally embedded vectors, involve many dimensions; each vector is projected in an N-dimensional space and the measure of semantic similarity considers all these dimensions to assess whether vectors are located close to or apart from each other. Each dimension explains some of the variance that distinguishes association patterns among all the words in a text corpus. It is hard, however, to interpret each dimension in practice, i.e., to give a univocal explanation about the reason why two vectors are close (or apart) when considering one specific dimension. One might therefore consider projecting these vectors on a limited subset of pre-defined dimensions and evaluate the position of each vector in the new space, to measure the association of a given word within a set of more narrowly defined underlying concepts.[18] We specify some classes that might be relevant for gauging the sentiments surrounding key Darwinian concepts over time: "goodness", "importance", and "morality". In order to create a given dimension, we have to define a list of words that express it. Following Jenkins (1958), who specified a list of terms for a variety of cultural dimensions, we consider pairs of antonym words related to the three areas that we want to measure. We then average the differences of all the vector pairs: $\frac{\sum_{P}^{|P|} \overline{p_1} - \overline{p_2}}{|P|}$, where $\overline{p_1}$ and $\overline{p_2}$ are the vectors of a one of P pairs of antonym words. The resulting vector represents an "average" dimension of the general underlying concept we aim to capture (e.g., Morality).

Finally, we calculate the similarity (or projection) of the vectors of key Darwinian words on each dimension and track the similarity over time. Figure 1B provides a simple graphical representation of how, for example, the word "Evolution" moves on the "Morality" spectrum. Each side includes a word that has an antonym on the opposite side (e.g., sinful, virtuous). The average of each pair forms a new dimension (represented by the underlying thick gray arrow in the figure) that should represent the concept of Morality in a comprehensive way. In practice, we measure the similarity between the word Evolution and the dimension spanned by $(\overrightarrow{good} - \overrightarrow{evil}) + (\overrightarrow{moral} - \overrightarrow{immoral}) + (\overrightarrow{virtuous} - \overrightarrow{sinful}) + \ldots$ By measuring the similarity by decade, we are able to assess how much this concept was deemed to be "moral" over time. The more positive a projection,[19] the stronger the association of Evolution with Morality.

[18] Kozlowski et al. (2019), for instance, project vectors related to different musical genres (e.g., jazz, rap, etc.) onto a plan that measures "affluence", to determine whether, say, jazz is more associated with wealthier strata of a population than rap, and how these associations vary over time.

[19] The projection of a word vector on a vector-dimension is equivalent to the cosine of the angle between the two vectors if the vectors are normalized.

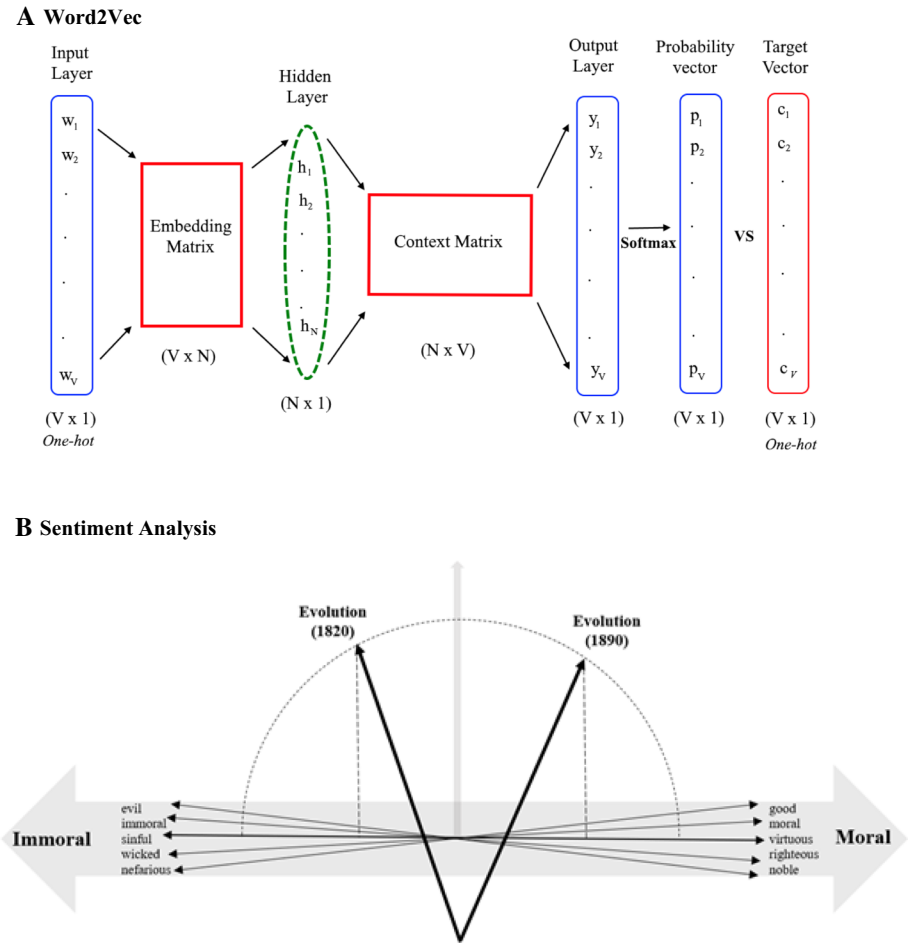


Fig. 1 Word2Vec model and Sentiment Analysis with Embedded Vectors. Panel A: Word2Vec. *Notes*: The diagram illustrates the structure of a Word2Vec model. Each word is encoded into binary vectors (one-hot) of dimension *V*x1. The embedding matrix (*V*x*N*) and the context matrix (*N*x*V*) are initialized with random weights (note that $N < V$). The multiplication of the initial one-hot vector and the embedding matrix gives us the embedding vector of the input word we are currently considering. This embedding vector forms a hidden layer of dimension *N*x1. The multiplication of the hidden layer and the context matrix forms the output vector, which becomes a probability vector after a soft-max transformation. This vector can be readily compared to the one-hot vector that identifies the considered context word (i.e., target vector). The difference between the probability and the target vector modifies the scores of the embedding and context matrix through a backpropagation mechanism so that the weight can be adjusted accordingly to real words co-occurrence. Panel B: Sentiment Analysis. *Notes*: the figure shows a subset of the pair of words that are used in the paper to span the "Morality" dimension. Two embedded vectors for the word Evolution for the periods 1820–29 and 1890–99 respectively are drawn and projected on the dimension. This figure exemplifies a change in perception that a word can go through over time

4 Findings

In the first part of this section, we analyze the evolution of the relative frequency of key words in Darwin's theory and expressions as measures of the diffusion of key concepts in

the public discourse around the time of the publication of *On the Origin of Species* in 1859. Then, we move to the analysis of semantic and sentiment changes concerning these words.

4.1 Frequency analysis

We consider terms (1-grams) that, from many accounts (Desmond & Moore, 1994, and Mayr, 1995), as well as our own reading, represent the key concepts in Darwin's theory: Evolution, Selection, Adaptation, Competition, Survival, and the expression (2-gram) Natural Selection.

We begin by computing the frequency of these words and expressions, overall and separately for fiction and non-fiction books, and in comparisons with other frequent nouns in *On the Origin of Species*. Second, we assess the diffusion of the main Darwinian concepts in languages different than English, in order to explore the diffusion of the theory of evolution in other cultures, and whether it happens with some delay. Third, we attempt to isolate the contribution of Darwin to the public discourse from the general "presence in the air" of ideas about evolution, by tracing the use of the word Darwin itself, as opposed to other scientists engaged in that field. Finally, we explore the diffusion of Darwin's ideas in the political debate.

4.1.1 Darwinian and "control" concepts; fiction and non-fiction books

Figure 2 reports the frequency of use of the key Darwinian terms Evolution, Selection, Adaptation, Competition, Survival, and the expression (2-gram) Natural Selection. The frequencies are per million words, in each year between 1820 and 1899, for the whole Google Book corpus (Panel A) and for non-fiction and fiction books separately (Panel B).

The expression Natural Selection, perhaps the most defining of Darwin's concepts, was virtually non-existent in both the fiction and non-fiction literature before 1859 and experienced a significant increase in the rate of adoption since then. On the one hand, this may not be surprising, precisely because of the close association of Darwin's work with the idea of natural selection. On the other hand, we may consider the significant increase in the diffusion of this concept immediately after the publication of *On the Origin of Species* as a validation of our approach; this initial analysis of frequencies does capture what we might have expected.

Evolution and Survival also substantially increased the adoption rate in the years immediately following the publication of Darwin's book. The ideas that underlie these words and expressions, therefore, generated interest in not only specialized or more educated circles, but plausibly also in the more general cultural context.[20] Moreover, the diffusion of these concepts in the fiction literature lagged the diffusion in the non-fiction literature by a few years. Competition was already present in the first part of the 19th Century, especially in the non-fiction literature, but experienced an increase in the adoption rate after 1860. Selection experienced a weaker increase in relative frequency around the publication of *On the Origin of Species*.[21] Although Selection was already present before 1859, Natural

[20] Interestingly, the increase in the use of Evolution occurred especially since the 1870s, coinciding with the use of the term in the 1872 edition of Darwin's book.

[21] In Figure A1 of the Appendix, we show additional frequency analyses where we "stem" the main Darwinian (Evolution, Selection, Survival, Competition and Adaptation) and consider also other words with the same roots. Specifically, we plot the frequency of one of five key words, and the average frequency of a set of other words with the same root. There are two main patterns. In the cases of Evolution, Competition

Selection, as an expression, appeared after the publication of *On the Origin of Species*. This suggests the possibility that, after 1859, the word Selection might have experienced a change of meaning and use in the public discourse. We investigate this below.

In Table 1, Slope(1820–59) and Slope(1860–99) are the parameter estimates from spline regressions of the frequency (per million words) of each of the Darwinian words and phrases on a time variable that represents each year between 1820 and 1899 (expressed as t = 20, 21, …, 99), with one knot at year 1859. Slope(1860–99)-Slope(1820–59) is the difference between the two slopes. Table 2 displays results from the same spline regressions, separately for fiction and non-fiction books. Finally, we ran spline regressions with knots at each decade between 1820 and 1899. The estimates are in Table 3; for a more parsimonious exposition, we aggregated the six Darwinian expressions into an index given by the average annual frequency. Word-by-word estimates are in the Appendix Table A2). Columns (1) through (3) display the estimates separately for fiction and non-fiction books, as well as overall.

The estimates reinforce the visual evidence in Figs. 2 and 3. They also confirm the delay in diffusion in the fiction literature that we observed in the graphical representations: the estimated slopes are large and statistically significant starting in the 1870s for the fiction subsample.[22] The extent of these changes is substantial. For example, the average frequency of the six Darwinian words and expressions oscillated between 5 and 10% of the standard deviation of the yearly frequency of all words present in the Ngram corpus in a given year between 1820 and 1859, and then began to rise up to 30% of the yearly standard deviation from 1860 to the end of the century (see Figure A2 in the Appendix).

In Column (4) of Table 3, we report estimates from of the average yearly frequency of a group of "control" or "placebo" words to compare to the terms that represent the key Darwinian concept. We selected the 100 most frequent nouns in *On the Origin of Species*, and then eliminated Selection, which is among these 100 nouns but also one of the Darwinian words. The remaining ninety-nine words are not specific to the theory of evolution. For these words, we do not detect any particular change in diffusion before and after 1859.[23]

Footnote 21 (continued)
and, to a lesser extent, Selection, the average frequency of the words with the same root appears to follow a similar pattern as the corresponding word of interest. However, these other words have substantially lower frequency and therefore presence in the written language; in other words, in these three cases our main word of interest within a given etymological root is dominant. In the other two instances, Survival and Adaptation, the words with those same roots have similar diffusion on average, but their pattern over time is erratic. This evidence corroborates our focus only on the key Darwinian terms rather than the stemmed words overall.

[22] The statistical significance of the estimates is robust to multiple hypothesis corrections; we applied the Romano-Wolf procedure to the six regressions whose estimates are in Table 1, to account for the use of multiple left-hand-side variables. See Romano and Wolf (2005a, b, 2016) as well as Clarke et al. (2019) for the Stata procedure. Standard error estimates are very similar if we use the Newey-West in lieu of the Huber-White correction (see Tables A8 through A10).

[23] In the Appendix, Table A1 reports the list of these words. The list includes both general terms like number, animals and nature, and more specific ones such as eggs and insects. In Appendix Figure A3, we display the relative occurrence of some of the ninety-nine control terms as an example: Nature, Number, Life, Animals, Flowers, Plants. For none of these words is there any discernible change in diffusion in the decades immediately preceding and following the publication of *On the Origin of Species*. Appendix Table A3 reports estimates from spline regressions (with one knot at year 1859) on these six words. Appendix Figure A4 plots the estimated frequency slope in 1820–59 and 1860–99 for each of the ninety-nine words; the estimates gravitate around the 45-degree line, thus indicating limited changes in the rate of diffusion after the publication of *On the Origin of Species*.

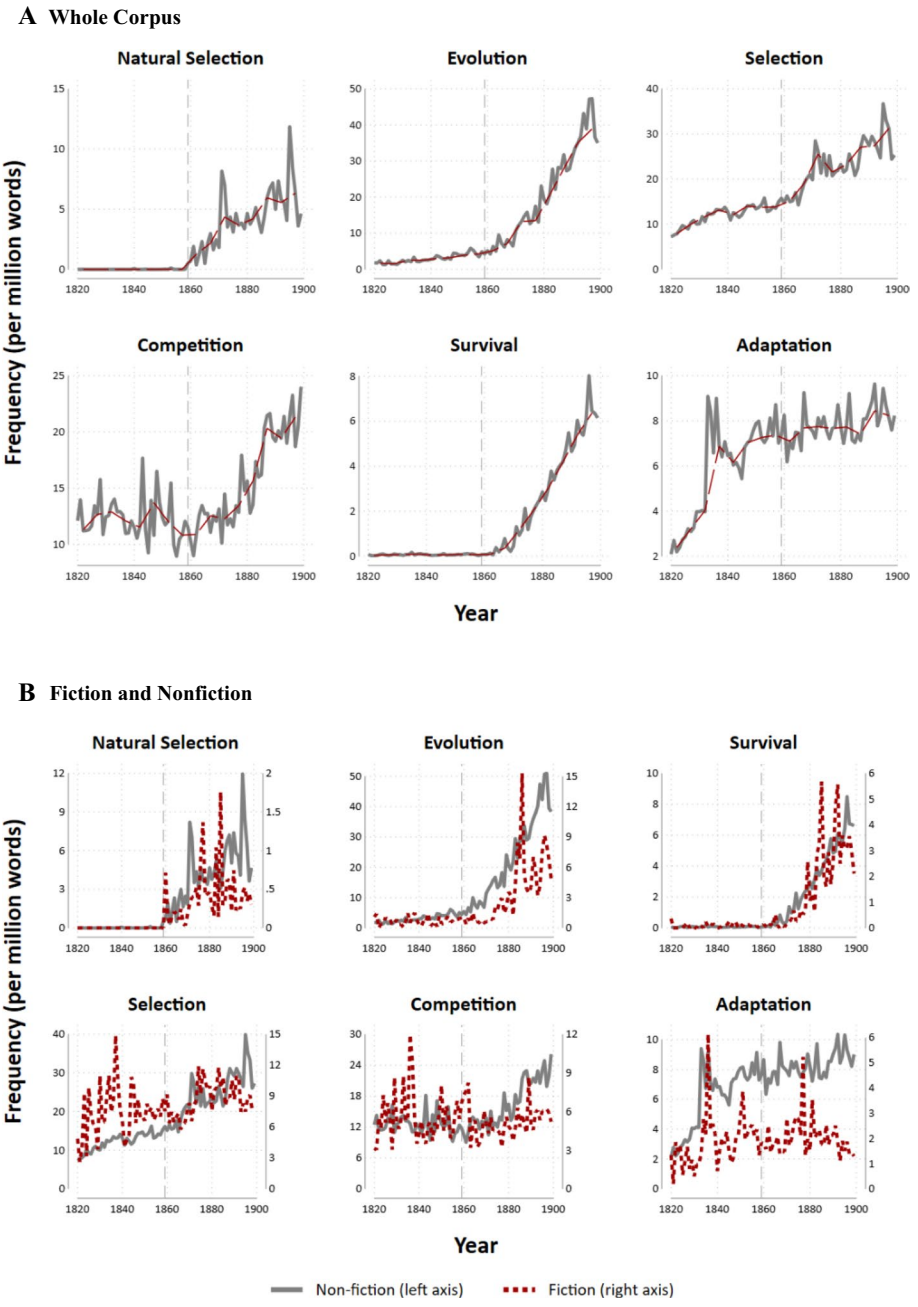


Fig. 2 Frequencies (per 1 Million Words) of Darwinian Concepts in the Google Books Corpora. *Notes*: For each year, the graphs show the number of occurrences of the word or phrase reported on top per one million words. In panel A, the gray solid line displays the yearly frequency, whereas the red dashed line is a median band plot with 16 intervals (each of 5 years). Note that also the denominators for the calculation of the relative frequencies are separate for fiction and non-fiction

Table 1 spline regression analyses (one knot) – frequency of Darwinian concepts

Word/phrase	Natural selection (1)	Evolution (2)	Selection (3)	Competition (4)	Survival (5)	Adaptation (6)
Slope (1820–59)	0.027***	−0.003	0.177***	−0.071***	−0.020***	0.177***
	(0.010)	(0.022)	(0.019)	(0.017)	(0.004)	(0.013)
Slope (1860–99)	0.165***	0.950***	0.372***	0.269***	0.171***	0.002
	(0.020)	(0.044)	(0.033)	(0.020)	(0.007)	(0.008)
Slope (1860–99)-Slope (1820–59)	0.138***	0.953***	0.194***	0.340***	0.191***	−0.115***
	(0.028)	(0.062)	(0.049)	(0.033)	(0.010)	(0.019)
Frequency (1820)	0.004	1.8	7.2	12.1	0.08	2.1
Frequency (1859)	0.5	4.3	15.8	11.5	0.07	7.4
Avg. frequency (1820–1859)	0.2	3.0	12.1	12.3	0.07	5.9
Avg. frequency (1860–1899)	4.3	21.2	23.3	15.8	2.9	7.8
Observations	80	80	80	80	80	80
R-squared	0.779	0.952	0.888	0.720	0.952	0.658

For each word and phrase, the two estimates Slope(1820–59) and Slope(1860–99) refer to the slopes of the best linear fit from a spline regression of frequency (per million words) on years from 1820 to 1899, expressed as t=20, 21, …, 99, with one knot at 59. The regression equation for a given word w is: $y_{wt} = \alpha_w + \beta_{1w}\min(t, 59) + \beta_{1w}(\max(t, 59) - 59) + \varepsilon_{wt}$. Slope(1860–99)-Slope(1820–59) represents the estimate of the difference between the two slopes. Robust standard errors are in parentheses

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

We further rely on this group of generic words to perform difference-in-difference analyses whose findings are in Table 4 and Fig. 3. In Table 4, we report the estimates from analyses where, for each year, we sum up the frequencies of the six Darwinian concepts on the one hand and of the ninety-nine control nouns on the other hand, and compare the trends in aggregate diffusion before and after 1859. Because the aggregate frequency of the generic words is much higher than the frequency of the Darwinian concepts pooled together, to make more immediate comparisons we transform these frequencies into their natural logarithms and include the logarithm of the time trend in the regression analyses. In this analysis, we also pool together fiction and non-fiction books. The regression model that we estimate is as follows:

$$\begin{aligned}\ln(y_{wt}) = {} & \alpha_w + \beta_w \ln(t) + \gamma_w(\ln(t) - \ln(59)) * 1(t > 59) + \delta_w 1(Darwinian) \\ & + \theta_w \ln(t) * 1(Darwinian) + \lambda_w(\ln(t) - \ln(59)) * 1(t > 59) \\ & + \mu_w(\ln(t) - \ln(59)) * 1(Darwinian) * 1(t > 59) + \varepsilon_{wt}\end{aligned} \tag{1}$$

The sample thus includes 160 observations, two for each year, with one reporting information about the generic words ($1(Darwinian) = 0$), and the other about the six Darwinian concepts ($1(Darwinian) = 1$). Columns (1) and (2) of Table 2 display estimates of a simplified version of the model, were the left-hand-side variable is the natural logarithm of the sum of frequencies of Darwinian and generic terms separately, regressed on a time trend and the interaction between the indicator for years greater than 1859 and the difference between the current year and 1859. Estimates of the parameters of the full model are in Column (3). The estimate of the coefficient on the interaction

Table 2 Spline regression analyses (one knot) – frequency of Darwinian concepts, separate for fiction and non-fiction books

Word/phrase	Natural selection		Evolution		Selection	
	Non-fiction	Fiction	Non-fiction	Fiction	Non-fiction	Fiction
	(1)	(2)	(3)	(4)	(5)	(6)
Slope (1820–59)	0.027**	0.005***	−0.009	−0.022***	0.183***	0.031
	(0.010)	(0.002)	(0.024)	(0.008)	(0.020)	(0.025)
Slope (1860–99)	0.167***	0.012***	1.028***	0.173***	0.412***	0.040**
	(0.020)	(0.003)	(0.049)	(0.020)	(0.036)	(0.016)
Slope (1860–99)-Slope (1820–59)	0.140***	0.007*	1.037***	0.195***	0.229***	0.009
	(0.028)	(0.004)	(0.068)	(0.25)	(0.052)	(0.034)
Frequency (1820)	0.004	0.0	1.9	1.1	7.3	4.9
Frequency (1859)	0.5	0.0	4.5	1.1	16.1	7.5
Avg. frequency (1820–1859)	0.02	0.001	3.1	0.7	12.3	7.4
Avg. frequency (1860–1899)	4.3	0.4	22.6	3.5	24.4	8.7
Observations	80	80	80	80	80	80
R-squared	0.779	0.397	0.949	0.629	0.890	0.145
Word/phrase	Competition		Survival		Adaptation	
	Non-fiction	Fiction	Non-fiction	Fiction	Non-fiction	Fiction
	(7)	(8)	(9)	(10)	(11)	(12)
Slope (1820–59)	−0.074***	−0.008	−0.021***	−0.013***	0.122***	0.025***
	(0.018)	(0.019)	(0.004)	(0.004)	(0.014)	(0.008)
Slope (1860–99)	0.300***	−0.000	0.178***	0.096***	0.010	−0.012**
	(0.022)	(0.014)	(0.007)	(0.009)	(0.009)	(0.006)
Slope (1860–99)-Slope (1820–59)	0.374***	0.007	0.199***	0.109***	−0.112***	−0.037***
	(0.036)	(0.029)	(0.011)	(0.012)	(0.020)	(0.012)
Frequency (1820)	12.4	3.2	0.1	0.4	2.1	1.3
Frequency (1859)	11.8	6.7	0.1	0.1	7.6	1.8
Avg. frequency (1820–1859)	12.6	5.6	0.1	0.1	6.1	1.8
Avg. frequency (1860–1899)	16.9	5.2	3.0	3.0	8.3	2.0
Observations	80	80	80	80	80	80
R-squared	0.740	0.004	0.950	0.725	0.673	0.088

For each word and phrase, the two estimates refer to the slope of the best linear fit from a spline regression of frequency (per million words) on years from 1820 to 1899, expressed as t = 20, 21, …, 99, with one knot at 59. The regression equation for a given word w is: $y_{wt} = \alpha_w + \beta_{1w}\min(t, 59) + \beta_{1w}(\max(t, 59) - 59) + \varepsilon_{wt}$. Slope(1860–99)-Slope(1820–59) represents the estimate of the difference between the two slopes. Robust standard errors are in parentheses

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

between the indicator for Darwinian words, the indicator for the post-1859 period and the difference between the current year and 1859 (μ_w) is positive, large and statistically significant, indicating a much larger relative increase in the frequency of Darwinian concepts after 1859. The estimate of θ_w is significantly smaller than the estimate of

Table 3 Spline regression analyses (eight knots) – average frequency of Darwinian and generic words

Word/phrase and sample:	Darwinian words: non-fiction	Darwinian words: fiction	Darwinian words: full sample	Generic words: full sample
	(1)	(2)	(3)	(4)
Slope(1820–29)	0.142***	0.104**	0.142***	0.990***
	(0.035)	(0.044)	(0.034)	(0.231)
Slope(1830–39)	0.073**	0.025	0.067**	0.333*
	(0.028)	(0.052)	(0.027)	(0.196)
Slope(1840–49)	0.065**	−0.049	0.060**	−0.185
	(0.029)	(0.052)	(0.028)	(0.160)
Slope(1850–59)	−0.031	0.044	−0.028	−0.042
	(0.036)	(0.027)	(0.034)	(0.156)
Slope(1860–69)	0.339***	−0.058**	0.312***	0.138
	(0.055)	(0.022)	(0.052)	(0.146)
Slope(1870–79)	0.213***	0.150***	0.219***	0.088
	(0.064)	(0.033)	(0.060)	(0.114)
Slope(1880–89)	0.565***	0.081**	0.512***	0.275**
	(0.057)	(0.040)	(0.054)	(0.115)
Slope(1890–99)	0.336**	−0.047	0.273*	−0.313
	(0.148)	(0.049)	(0.138)	(0.190)
Avg. frequency (1820–29)	4.44	2.16	4.352	148.99
Avg. frequency (1850–59)	6.49	2.75	6.30	157.03
Avg. frequency (1890–99)	19.39	4.31	18.10	158.98
Observations	80	80	80	80
R-squared	0.967	0.637	0.966	0.588

Each yearly observation is the average frequency (per million words) of the six main Darwinian words and phrases (full sample as well as separate between fiction and non-fiction) and of the 99 most frequent nouns in *On the Origin of Species* that we use as our control group of generic words. The estimates refer to the slope of the best linear fit from a spline regression of frequency (per million words) on years from 1820 to 1899, expressed as t = 20, 21, …, 99, with eight knots at 19, 29, …, 89

The regression equation for a given word w is: $y_{wt} = \alpha_w + \beta_{1w}\min\{t, 29\} + \beta_{2w}(Max\{\min\{t, 39\}, 29\} - 29) + \beta_{3w}(Max\{\min\{t, 49\}, 39\} - 39) + \beta_{4w}(Max\{\min\{t, 59\}, 49\} - 49) + \beta_{5w}(Max\{\min\{t, 69\}, 59\} - 99) + \beta_{6w}(Max\{\min\{t, 79\}, 69\} - 69) + \beta_{7w}(Max\{\min\{t, 89\}, 79\} - 79) + \beta_{8w}(Max\{\min\{t, 99\}, 89\} - 89) + \varepsilon_{wt}$

Robust standard errors are in parentheses

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

μ_w, but it is positive and statistically different from zero; this indicates that also before 1859, the frequency of Darwinian concepts was increasing at a higher rate that the combined generic terms. This is likely due to the trend and diffusion that some Darwinian terms, such as Selection and Adaptation, were experiencing also in the first half of the 19th Century. The trend, however, clearly had an additional, substantial acceleration after the publication of *On the Origin of Species*.

Second, we define a model where the outcome variable is the annual frequency (from 1820 to 1899) of each of the six Darwinian concepts and of the ninety-nine control nouns separately. Here we estimate the average difference in frequency for the Darwinian words and the generic words in each decade:

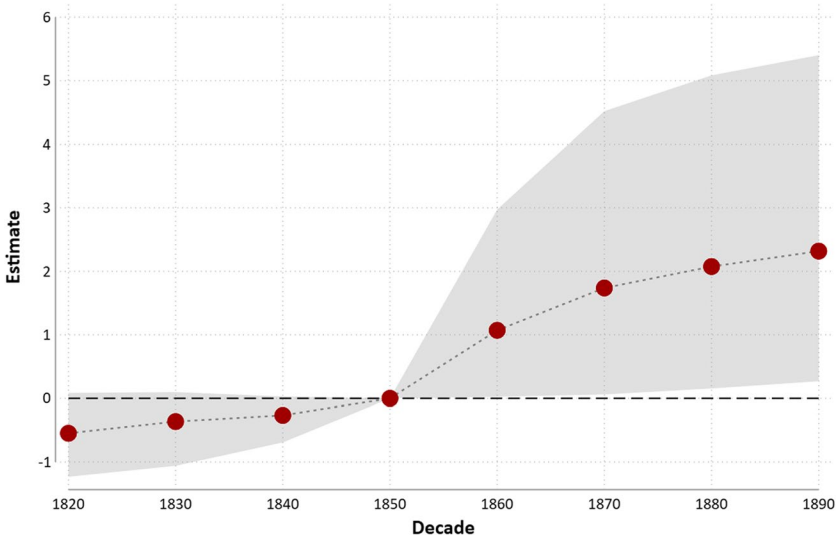


Fig. 3 Differences-in-Differences Estimates of the Average Frequency of Darwinian and Generic Concepts in Each Decade between 1820 and 1899. *Notes*: Each dot in the graph represents the estimate of the parameters δ_j from the following regression model: $\ln(y_{wt}) = \alpha_w + \beta_w 1(Darwinian) + \sum_{j=2}^{4}\gamma_i 1(j0 \le t \le j9) + \sum_{j=6}^{9}\gamma_j 1(j0 \le t \le j9) + \sum_{i=2}^{4}\delta_i 1(j0 \le t \le j9) * 1(Darwinian) + \sum_{j=6}^{9}\delta_j 1(j0 \le t \le j9) * 1(Darwinian) + \varepsilon_{wt}$, where y_{wt} is the frequency of use of a word per million words used. Each value on the x-axis correspond to a decade; for example, 1830 corresponds to 1830–39. The omitted (or baseline) decade is 1850–59 (1850). Because the observed frequency is equal to zero in some cases, we add 0.001 to each frequency (0.001 is half of the lowest positive frequency per million words in our sample). The shaded area represents 95% confidence intervals that we computed using a wild bootstrap procedure (Roodman et al., 2019). Results are almost identical if we use an arcsine: $z = \ln\left(y + \sqrt{y^2 + 1}\right)$, or if we apply the GMM procedure described in Bellego and Pape (2019) to estimate the parameters (the confidence intervals from bootstrapped standard errors are narrower in this last case).

$$\begin{aligned}\ln(y_{wt}) = {} & \alpha_w + \beta_w 1(Darwinian) + \sum_{j=2}^{4}\gamma_i 1(j0 \le t \le j9) \\ & + \sum_{j=6}^{9}\gamma_j 1(j0 \le t \le j9) + \sum_{i=2}^{4}\delta_i 1(j0 \le t \le j9) * 1(Darwinian) \\ & + \sum_{j=6}^{9}\delta_j 1(j0 \le t \le j9) * 1(Darwinian) + \varepsilon_{wt}\end{aligned} \tag{2}$$

This analysis is on (6 + 99)*80 = 8,400 observations. The omitted time category is the decade 1850–59 ($50 \le t \le 59$). The δ_j coefficients thus indicate the difference between Darwinian and control terms, as compared to the reference difference in the 1850–59 period. Figure 3 displays the estimates of the δ_j coefficients and their 95% confidence intervals, and shows that the difference in relative frequency between the Darwinian and generic terms is much larger after the publication of *On the Origin of Species* than before. The main unit of observation in this analysis is a given term, so we cluster standard errors at that level. We have, therefore, 105 clusters; however, the number of "treated" units (and consequently the number of treated clusters) is small relative to the control ones. As such, to estimate confidence intervals around each of the estimated difference-in-differences parameters, we rely on the subcluster wild bootstrap procedure of MacKinnon and Webb (2018; see also

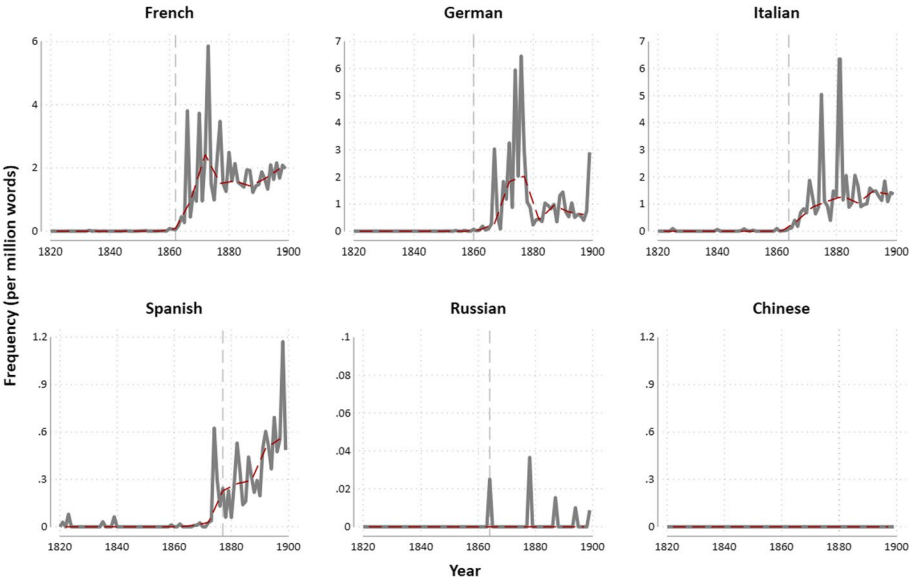


Fig. 4 Frequencies (per 1 Million Words) of the Phrase "Natural Selection" in Six Languages Other than English. *Notes*: For each year, the figures report the number of occurrences (per million words) of the expression "Natural Selection" in the language indicated on top of a graph. The red dashed lines are a median band plot with 16 intervals (each of 5 years). The vertical dashed line are in correspondence of the year of the first published translation of *On the Origin of Species* in a given language

Roodman et al., 2019). Despite the larger estimated confidence intervals, the estimates still show a significant change after 1859.

To assess the robustness of these last two analyses, we also defined a second control group; we selected the 100 words whose frequency between 1855 and 1858, the years immediately before the publication of Darwin's book, was closer in absolute value to the average frequency of the six Darwinian expressions. Whereas the selection of first group of words was motivated by the fact that those terms were present in *On the Origin of Species*, the rationale for this second group is the similar diffusion in the public discourse. Appendix Figure A5 in the Appendix presents the same type of plot as the one in Fig. 3, from a regression with the alternative control group; the patterns are remarkably similar.[24]

In addition to the comparison with two sets of words, we address a further concern that the significant change in the frequency of Darwinian terms after 1859 may be due to an overall change in the composition of texts, at least in the Ngram corpus. Although this corpus does not identify books, but only words and expressions, we can assess whether the total number of words, and the rate of "entry" and "exit" of words in the corpus, was different in the years around the publication of On the Origin of Species as compared to other periods. Figure A6 and Table A5 in the Appendix shows that his was not the case.

[24] The list of these alternative control words is in Appendix Table A4.

Table 4 Differences-in-differences regressions – Darwinian and generic scientific concepts

Outcome variable:	ln(aggregate frequency)		
Sample:	Generic words	Darwinian words	Darwinian and generic words
	(1)	(2)	(3)
Regressors			
ln(Year)	0.042***	0.405***	0.042***
	(0.010)	(0.035)	(0.010)
1 (Year > 1859) x ((ln(Year)-ln(59))	−0.021	1.723***	−0.021
	(0.020)	(0.092)	(0.020)
1(Darwinian word)			−4.668***
			(0.134)
1(Darwinian word) × ln (year)			0.363***
			(0.036)
1(Darwinian word) × 1 (year > 1859) × ((ln (year)-ln(59))			1.744***
			(0.095)
Constant	9.482***	2.011***	4.886***
	(0.038)	(0.129)	(0.038)
Observations	80	80	160
R-squared	0.407	0.966	0.998

Columns 1 and 2 report estimates from regressions where the outcome variable is the natural logarithm of the aggregate frequency of the 99 most frequent nouns in *On the Origin of* Species (column 1) and of the aggregate yearly frequencies of the six Darwinian word and concepts (column 2). The regression estimates in column 3 come from combining the data used for the regressions in columns 1 and 2; therefore there are two observations per year (N = 160), with one reporting information about the generic words ($1(Darwinian) = 0$), and the other about the six Darwinian concepts ($1(Darwinian) = 1$). The regression equation is: $\ln(y_{wt}) = \alpha_w + \beta_w \ln(t) + \gamma_w(\ln(t) - \ln(59)) * 1(t > 59) + \delta_w 1(Darwinian) + \theta_w \ln(t) * 1(Darwinian) + \lambda_w(\ln(t) - \ln(59)) * 1(t > 59) + \mu_w(\ln(t) - \ln(59)) * 1(Darwinian) * 1(t > 59) + \varepsilon_{wt}$

4.1.2 Translation in other languages

Were the effects of *On the Origin of Species* specific to the social context in which the book was written and first published? Or did the treatise generate a similar impact in other countries upon its translation? Moreover, did the diffusion of scientific concepts in the cultural environment relate to the status of a country economic development, literacy rate, or development of the publishing industry? To answer these questions, we study whether the translations of *On the Origin of Species* generated a similar the diffusion of its key concepts in other languages.

As shown in Fig. 4, the phrase Natural Selection substantially increased its diffusion upon the translation of *On the Origin of Species*. The same holds for such words as Evolution, Survival, and Competition (Figure A7 in the Appendix). Moreover, the frequency of use of most words started increasing right after 1859, indicating that, even in the absence of an official translation, Darwin's concepts diffused across borders. These results suggest that the cultural effects of *On the Origin of Species* were not specific to the English-speaking context.

We cannot claim that the translation years are exogenous. For example, the translation might have occurred first in countries where the interest was higher, and this, in turn, might have affected diffusion. The likely endogeneity of the publication year, however, offers an opportunity for additional considerations about the relationship between the broad cultural acceptance of scientific concepts and economic development. For instance, in countries like Italy and Spain, both "late comers" during the Industrial Revolution (Ciccarelli & Nuvolari, 2015), the translation of *On the Origin of Species* occurred later than the translation into French and German, i.e., the languages of two countries where industrialization occurred earlier. Conclusions for Russian and Chinese terms are more tentative, because the N-gram repository plausibly includes a relatively small number of books in these languages. Nonetheless, Russia was mostly a feudal country until World War I (Markevich & Zhuravskaya, 2018), and had the first translation of Darwin's book even later; and China was long isolated from the scientific debate, which, according for example to Mokyr (2008), delayed its industrial development. It is perhaps not surprising, given the features of these two countries, that the diffusion of Darwinian words and phrases was extremely limited in Russian and Chinese books in the 19th Century.[25]

The translation year may not only depend on the status of a country economic development, but also on the literacy rate or the development of the publishing and translating industries. We collected data on these variables from the countries' Censuses between 1800 and 1950 for the available years. Darwin's concepts diffused first in countries with a higher literacy rate (like the UK, US, France, and Germany) than in countries with a low literacy rate (such as Italy, Spain and Russia). Similarly, where the publishing or translating industries were more developed, Darwin's idea diffused sooner. This is the case of France,

[25] With the exception of Chinese, the other languages that we considered are linked to a predominantly Christian culture, where creationism was well accepted. The N-gram database does not include many languages that refer to non-Christian environment. In addition to Chinese, the only available one is Hebrew. We report our analyses on this language in Appendix Figure A8. Because a Hebrew version of *On the Origin of Species* was only available starting from 1960, we also extended the period of observation to the end of the XX century to 2000 in order to assess the evolution of Darwinian words after the book's publication. We observe minimal, if any, use of most of the Darwinian terms in the XIX century, with an increase in diffusion in the second half of the 1900s.

and Germany. By contrast, countries with less diffusion also experienced a lower development of the publishing industries, like Spain, Italy, Russia, and China.

4.1.3 Ideas in the air, substitution and multiple attribution

The various findings that we just reported show that some concepts, as measured by the words that embody them, were only marginally present in the public discourse before the publication of *On the Origin of Species*. However, even if they entered the public discourse only after 1859, certain terms may have simply substituted existing ones while expressing the same ideas. In Fig. 5 we report the frequency of occurrence of the names of four scientists who contributed, in different ways, to the understanding of evolution. In addition to Darwin, we consider Alfred Russell Wallace, Robert Chambers, and Jean-Baptiste Lamarck. Lamarck's theory of the transmission of acquired traits is frequently mentioned as an example of "failed" theory to compare to Darwin's. Chambers' *Vestiges of the Natural History of Creation* introduced, in the 1840s, the idea of an "evolution" of living and non-living beings over time, more as a speculation than as a complete scientific treatment (note that the author was anonymous until 1884). Alfred Russell Wallace's work was close, in time and content, to Darwin's. Figure 5 shows that both Darwin and Wallace increased their occurrence in the English book corpus in the second half of the 19th Century, but Darwin's frequency increased substantially more. Chambers and Lamarck were already present before then, but their frequency remained stable (and low) after 1860.[26] The estimated difference in the increase of diffusion after 1989 between Darwin and the other names are statistically significant (see also Appendix Table A6).

Because Lamarck was (and originally wrote in) French, we then compare the diffusion of the words Darwin and Lamarck in the French corpus. After 1860, the relative occurrence of the word Darwin in French books surpassed the frequency of Lamarck. We also compare terms that related to the study of the emergence and development of new species: Evolution and Transmutation. Although Evolution, which we already analyzed above, is typically associated with Darwin's work, earlier works in biology (including some of Darwin's) used the term Transmutation to characterize (gradual or discrete) transformations of plants and animals. By comparing these two words, we want to assess whether the broader literature and cultural discourse also picked up the "newer" word to express these changes. For books in French, we consider the word Transformism (Transformisme in French), which was used by Lamarck. The graphical representation of our findings is in Fig. 6. The general pattern is that Evolution became progressively more frequent than Transmutation, with a significant change in frequency after the 1850s. The substantially larger frequency of Evolution also suggests that this word did not just "replace" words that expressed overall similar concepts, but plausibly represent a broader diffusion of certain new ideas.

Overall, this evidence suggests that Darwin, with his own work and especially his 1859 book, caused a discontinuous change in the cultural discourse.

[26] We add the following combinations of names, middle names and last names: Alfred Russel Wallace, Alfred Wallace, Charles Darwin, Charles Robert Darwin, Robert Chambers, Jean-Baptiste Lamarck, Jean-Baptiste de Lamarck, Jean Baptiste Lamarck, Jean Baptiste de Lamarck.

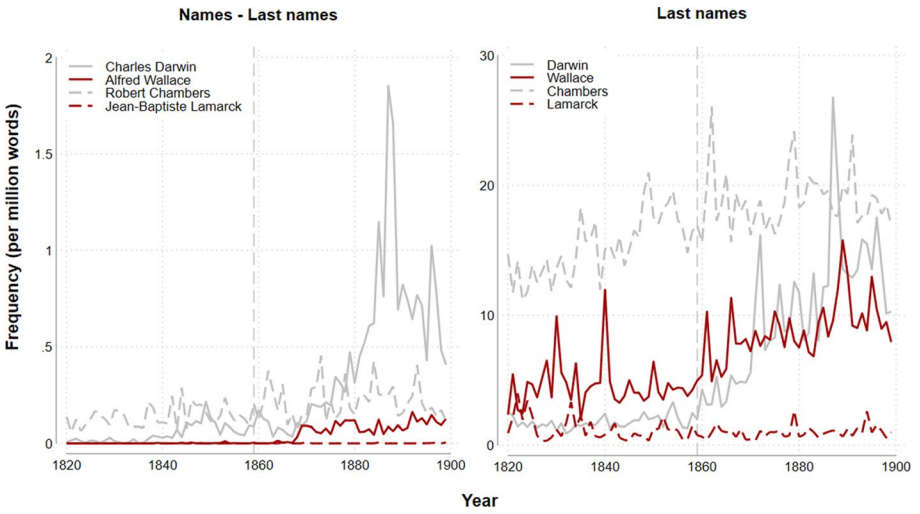


Fig. 5 Frequencies (per 1 Million Words) of Occurrences of the Names Charles Darwin, Alfred Wallace, Robert Chambers and Jean-Baptiste Lamarck in the English Google Books Corpus. *Notes*: For each year, the figures report the number of occurrences (per million words) of the name indicated in the legend. When we consider both the first and last names (left panel), we include different combinations of the full names of the four scientists: Alfred Russel Wallace, Alfred Wallace, Charles Darwin, Charles Robert Darwin, Robert Chambers, Jean-Baptiste Lamarck, Jean-Baptiste de Lamarck, Jean Baptiste Lamarck, and Jean Baptiste de Lamarck. The vertical dashed line is in correspondence of the year of the year of publication of *On the Origin of Species* (1859)

4.1.4 The diffusion of Darwinian concepts in the political arena

We perform frequency analyses on the UK Parliamentary debates and U.S. Congress data to assess whether Darwin's theory spilled over not only to the cultural discourse, but also to the political debate; arguably, culturally accepted scientific concepts may also affect how laws are shaped.

The UK Parliamentary data include a transcription of the debates in the House of Lords and the House of Commons. The corpus of Congressional Records includes the transcripts of all legislative debates occurring on the floor of the US Congress. It also contains additional materials, such as communications from the president and the executive branch agencies memorials, petitions, and supplementary information on the current legislation. We argue that these two corpora represent the official and most comprehensive daily account of the political discussion happening in the United Kingdom and United States. Although the text corpora of parliamentary debates are smaller than those we used on the main analysis, we think they can still offer suggestive evidence of the diffusion of the Darwinian concepts in the political debate.

Figure 7 shows an increase of the frequencies of such words and concept as Evolution, Survival and Natural Selection in both the Parliamentary and Congress debates after 1859. The evidence of an increase in use of these terms is clearer for the US Congress than for the UK Parliament. Overall, these results suggest that, after diffusing in the cultural environment, key Darwin's concepts also reached the political debate. The ten-year median bands, in particular, indicate a change in the use of these words a few years after we see these changes in the Google Books data. The lag may suggest that the cultural diffusion

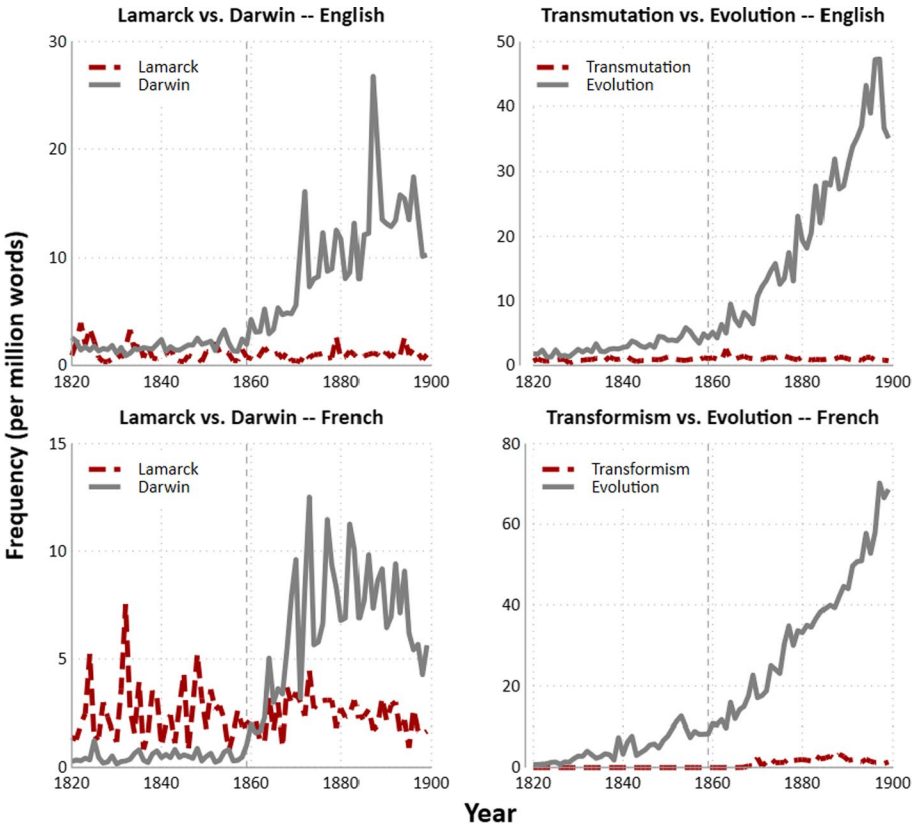


Fig. 6 Frequency of Occurrence (per million words) of the Words Darwin, Lamarck, Transmutation, Transformism and Evolution in the English and French Google Book Corpora. *Notes*: For each year, the figures report the number of occurrences (per million words) of the name indicated in the legend. The vertical dashed lines are in correspondence of the year of the year of publication of *On the Origin of Species* (1859)

was faster than, and perhaps a pre-condition for political diffusion.[27] In the case of the US, the Civil War in the first half of the 1860s may have further delayed the introduction of these new concepts in the legislative debate (Masci, 2019). Another explanation for the delay we observe might be that prior to 1837 each House was only required to keep an internal journal of its proceedings. External reporters could report verbatim debates only after that year. This might have hindered our capacity to fully capture the presence of Darwinian concepts in the initial period of the analysis.

[27] Spline regression analyses confirm that the words more related to Darwin's theory were significantly more likely to enter the political debate, after 1859, especially in the US Congress, with some lag with respect to the publication of On the Origins of Species. The less significant estimates in the UK Parliament corpus may be due to the fact that some of the Darwinian terms were perhaps more common in the UK than in other English-speaking countries such as the US. We do not find any specific pattern for the most frequent words in Darwin's book (estimates are in Appendix Table A7).

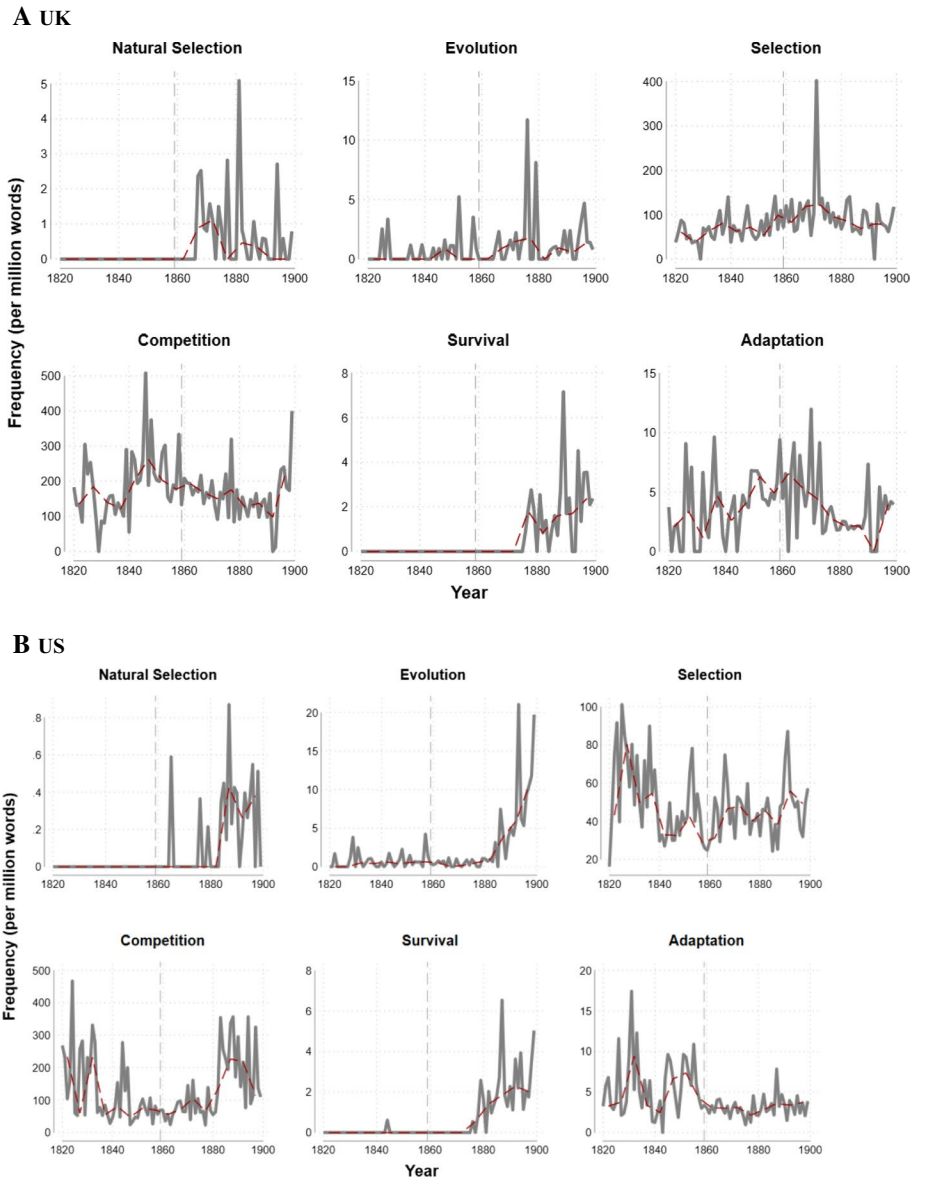


Fig. 7 Frequencies (per 1 Million Words) of Selected Darwinian Words and Phrases in the UK Parliamentary Debates and US Congressional Records. *Notes*: For each year, the graphs show the number of occurrences of the word or phrase reported on top per one million words. The gray solid line displays the yearly frequency, whereas the red dashed line is a median band plot with 16 intervals (each of 5 years)

4.1.5 Divine versus Darwinian creation

Although, according to many accounts, Darwin did not intend his theory of evolution through natural selection to go against religious (Christian) beliefs and doctrine,

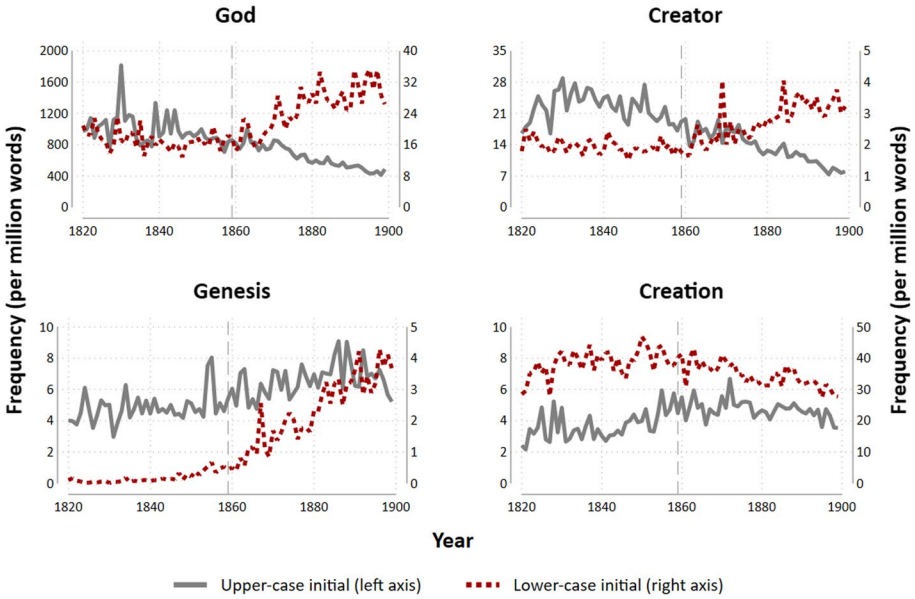


Fig. 8 Creation and the Theory of Evolution: Frequencies of Words and lower and upper case initials. *Notes*: For each year, the graphs show the number of occurrences of the word or phrase reported on top per one million words. The gray solid line displays the yearly frequency of the version of the word with an upper-case initial, whereas the red dashed line shows the frequency of the version starting with a lower-case letter

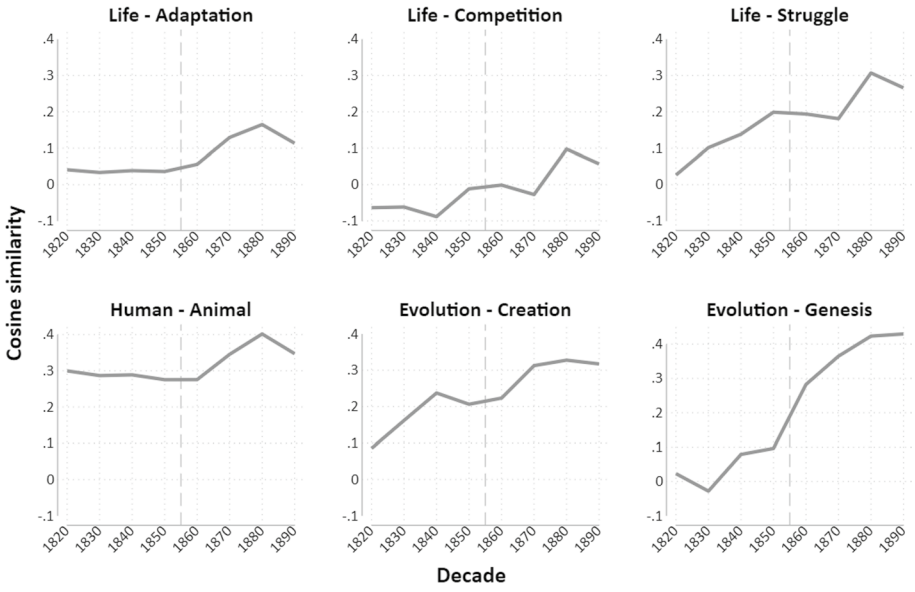


Fig. 9 Semantic Associations between Selected Pairs of Words. *Notes*: The graphs report the similarity between each pair of words, as measured by the cosine of the angle between each pair of word vectors. The weights in the word vectors were calculated with a Word2Vec algorithm. On the x-axis, 1820 represents the decade 1820–29, 1830 represents the decade 1830–39, and so on

implications of his discoveries such as the common origins of species, random variation and the absence of an intelligent design were largely perceived as a major blow to the Christian view of creation. In addition to exploring the diffusion of Darwinian concepts into the political discourse, a further way to assess how influential the cultural diffusion of the theory of evolution was is to investigate how certain topics that concerned both the religious sphere and Darwin's investigation evolved over time.

We next investigate if Darwin's theory had any impact on religion by focusing on specific terms with a strong religious root but also related to Darwin's theory. We analyze two terms related to the origins of the world, Creation and Genesis, the world Creator, which is one of the characterization of God in many religions, and the word God itself. We take advantage of certain rules or conventions in written text, when certain words are used in a religious context: the expression of the initial letter in upper case. In Fig. 8, we report the yearly frequency of use of the words God, Creator, Creation and Genesis with and without an upper-case initial. The increase in the use of the lower-case version of God, Genesis and Creator is visibly faster than the upper-case equivalent, perhaps indicating an overall process of relative "secularization" of the cultural domain. More relevant for our analysis, we observe a change in growth rate for the lower-case version of these three words again around 1860, whereas the upper-case equivalent terms follow a trend that does not change meaningfully for the whole eighty-year period around the publication of *On the Origin of Species*. For the word Creation, plausibly a term with a broader set of uses and meanings than the other three, there is no particular pattern for either the lower-case or the upper-case version. Overall, we interpret this evidence as showing that certain terms and underlying ideas with a strong religious connotation became more relevant also in the non-religious discourse. Below we report our explorations of semantic change where we will also assess the evolution of the use of certain terms with religious connotation by investigating changes in their meaning.

4.2 Semantic and sentiment analysis

Word embedding techniques require very large sample sizes to produce reliable results and insights. For this reason, in this section we limit the analysis to the Google Book database, and aggregate the data at the decade level.

4.2.1 Semantic analysis

Figure 9 introduces the second part of our study, where we move from the analysis of the frequency of use of certain words and the concepts underlying them, to the analysis of the semantic evolution of certain words and concepts, to see whether this evolution occurred in ways that we can relate to Darwin's theory. In the graphs, the horizontal axis reports decades (the time unit of reference), and the vertical axis indicates the cosine between the two-word vectors of interest.

One aspect of Darwin's theory is that life (or existence) includes adaptation, as well as competition, among its defining aspects. There is an increase in the semantic association between Life on the one hand, and Adaptation, Struggle and Competition on the other hand, especially after 1859. For Life and Struggle, we see a trend since the early 19th Century. We also investigate themes that presumably represented a controversy with the religious approach to the origins of species. One implication of Darwin's theory is that evolution applies to humans in the same way as it applies to other animals; although Darwin did not explicitly treat the human species

Table 5 Top 10 most similar words for selected Darwinian words

Decade	1820–29	1830–39	1840–49	1850–59	1860–69	1870–79	1880–89	1890–99
Adaptation	Structure	Fitness	Component	Fitness	Structure	Environment	Environment	Environment
	Configuration	Structure	Mechanism	Mechanism	Requirements	Structure	Conducing	Multiformity
	Durability	Nature	Fitness	Artistic	Assimilation	Differentiation	Organism	Exemplifies
	Fitness	Admirable	Arrangement	Emotional	Fitness	Reproduction	Adjustments	Adjustments
	Complexity	Relation	Nature	Agreeableness	Reproduction	Fitness	Aptitudes	Complexity
	Mechanism	Capabilities	Structure	Conducing	Exigencies	Assimilation	Selective	Functioning
	Harmonize	Causality	Deductive	Modulation	Structures	Organism	Complexities	Organism
	Nature	External	Conformation	Accomplishes	Conditions	Adjustments	Textures	Functionally
	Unfitness	Analogies	Suited	Phenomenal	Copiousness	Modification	Simplification	Reproduction
	Congruity	Mechanism	Optical	Structure	Arrangement	Requirement	Correlations	Definiteness
Evolution	Caloric	Sulphurous	Disengagement	Disengagement	Phenomena	Phenomena	Segregation	Integration
	Sulphuretted	Condensation	Lactic	Decomposition	Formation	Equilibration	Dissociation	Differentiation
	Decomposition	Oxidation	Acid	Combustion	Organic	Organic	Phenomena	Multiformity
	Absorption	Combustion	Decomposition	Absorption	Organism	Differentiation	Integration	Theory
	Combustion	Hydrogen	Nitrous	Carbonic	Integration	Theory	Divergences	Organic
	Carbonic	Absorption	Gas	Undulatory	Differentiation	Hypothesis	Process	Genesis
	Nitrous	Oxygen	Carbonic	Formation	Organisms	Organisms	Darwinian	Integrations
	Phosphorus	Germination	Hydrogen	Acid	Decomposition	Organism	Anhydride	Stages
	Chlorine	Gas	Decomposed	Metamorphosis	Molecular	Heterogeneity	Definable	Heredity
	Gases	Atomic	Oxygen	Fermentation	Absorption	Transformation	Differentiation	Functioning

Table 5 (continued)

Decade	1820–29	1830–39	1840–49	1850–59	1860–69	1870–79	1880–89	1890–99
Selection	Arrangement	Judicious	Judicious	Judicious	Collection	Sexual	Natural	Natural
	Judicious	Compilation	Arrangement	Arrangement	Choosing	Adaptation	Heredity	Heredity
	Transposition	Arrangement	Collection	Collection	Variation	Collection	Variation	Fittest
	Discrimination	Discrimination	Suitable	Recipes	Arrangement	Adaptations	Modification	Judicious
	Specification	Collection	Combination	Suitable	Adaptation	Preservation	Methodically	Sexual
	Proper	Choosing	Fitness	Translations	Suitable	Variation	Epigrammatic	Adaptation
	Illustration	Careful	Choosing	Proper	Divergence	Natural	Darwinian	Suitable
	Collection	Unpublished	Adaptation	Indexes	Variations	Judicious	Fittest	Variation
	Gleaned	Adaptation	Variety	Drawings	Survival	Instinct	Judicious	Choosing
	Careful	Models	Fittest	Careful	Discrimination	Heredity	Supplemented	Superintend
Survival					Geologic	Fittest	Fittest	Fittest
					Fittest	Copernican	Primitive	Evolution
					Minorities	Antipathy	Monogamy	Existence
					Equilibration	Evolution	Evolution	Outcome
					Reproducing	Eliciting	Darwinian	Result
					Necessitates	Extinction	Inferable	Modification
					Propulsion	Perpetuation	Chieftainship	Militancy
					Derangements	Existence	Conducing	Struggle
					Correspondences	Theism	Perpetuation	Subserves
					Computations	Rudiment	Preservation	Persistence

Table 5 (continued)

Decade	1820–29	1830–39	1840–49	1850–59	1860–69	1870–79	1880–89	1890–99
Competition	Emulation	Producers	Producers	Unrestricted	Producers	Markets	Producers	Producers
	Rivalship	Manufacturers	Market	Rivals	Commodities	Producers	Markets	Monopoly
	Markets	Market	Commodities	Producer	Prices	Monopoly	Monopoly	Capitalists
	Manufactures	Rivalship	Producer	Market	Dealers	Market	Market	Markets
	Rivals	Commodity	Monopoly	Markets	Consumer	Contend	Rivals	Overstocked
	Buyers	Collision	Demand	Gainers	Monopoly	Commodities	Trade	Unrestricted
	Market	Manufacturer	Lowers	Producers	Profits	Manufacturer	Employment	Conflict
	Trade	Emulation	Manufacturer	Grower	Markets	Trade	Commodities	Struggle
	Artisans	Sellers	Prices	Collision	Capitalists	Profits	Disadvantage	Traders
	Traders	Employment	Trade	Holders	Consumers	Stimulus	Buyers	Consumers

The table reports the ten most similar words (by cosine similarity) for each Darwinian word (Adaptation, Competition, Evolution, Selection, Survival) and each decade. We excluded synonyms and words with the same root from the graphs

in his 1859 book, this was the topic of his 1871 *The Descent of Man and Selection in Relation to Sex*. The semantic evolution of the word Human shows an increase in its similarity with Animal especially in the late 1800s. Furthermore, we investigate whether Darwinian concepts at the basis of the process of the birth of species, in particular Evolution, came to relate with terms that expressed this process in the religious discourse: Creation and, even more specifically, Genesis. The last two panels of Fig. 9 show a growing sematic similarity of Creation and Genesi with Evolution through the 19th century, consistent with a "secularization" of the discourse about the origin of the world. Again, the change in semantic similarity seems to accelerate starting in the 1860–70 decade.[28]

A second analysis of semantic changes focuses, again, on the key words and concepts that we considered so far. However, instead of investigating the similarity of these words with a select sample of other concepts, we "let the data speak" by determining, for each decade, the ten words with the highest semantic connection (cosine similarity) to these key words.[29] Table 5 reports the findings. We excluded from the rankings the words that had the same root as the focal key word as well as the most obvious synonyms (e.g., Compete or Competitor for Competition); we also defined a lower bound to the relevant cosine similarity to be equal to 0.05.

The table identifies a few interesting facts. First, the term Adaptation became, over the 19th Century, less related to physical or "mechanical" terms (such as Mechanism) and increasingly similar to concepts that represented living beings (such as Organism and Reproduction).

Second, substantial changes in meaning and association concern the word Evolution. In the first half of the 19th Century, the terms that were closest to Evolution came mostly from chemistry and physics. Later in the 1800s concepts from biology as well as related to human society were semantically more similar to Evolution. Examples include Social and Progress. Note also how the word Darwinian itself became closely associated with Evolution.

Third, Selection was more closely related to the concept of Choice (and qualification for the choice such as "careful" or judicious") in the first half of 1800; the similarity in meaning with Choice remained also later, but Selection also became more similar in meaning to other specific "Darwinian" words, such as Survival, Variation, Fittest and Heredity.

Fourth, very few words had a similarity in meaning with Survival, likely because the word itself was only rarely used in the first half of the 19th Century. Later in the century, the word was increasingly associated to other concepts related to evolutionary theory, notably Fittest, Evolution, Struggle and Selection. The increasing relatedness with Fittest toward the end of the 1880s is likely due also to the publication of the *Principles of Biology* by Herbert Spencer in 1864, where this concept applies also to society and ethics and

[28] In Figure A9 of the Appendix, we report additional analysis of semantic similarities between pairs of words. The graphs show, on the one hand, that some of the patterns in Fig. 9 are not specific to a narrowly defined pair of words. For example, the similarity pattern over time between Human and Animal is similar to the one for Human and Ape, Man and Animal, and Man and Ape. On the other hand, broader semantic trends that indicate the role of science in society, and not specific aspects of the theory of evolution, do not experience any change around 1860 (see for example Science-Knowledge, Religion-Knowledge, Science-Religion, Science-Nature, Evolution-Nature).

[29] These two approaches are similar to those proposed in Hamilton et al., (2016a, 2016b).

not only to the natural sphere. Competition, in contrast, maintained an association with a stable set of words, mostly related to production and markets, throughout the century.[30]

In Fig. 10 we display the semantic association between the five key words in *On the Origin of Species* that we took as expressing Darwin' contribution (Evolution, Selection, Survival, Competition and Adaptation), and the names of the four scientists (including Darwin) we considered in Sect. 4.1.3 above. With this exercise, we explore whether these key terms that defined the theory of evolution by natural selection were, in fact, specifically associated with Darwin or were part of a discourse that also included the contribution of other scientists. In general, the similarity of these words with Darwin is systematically positive and greater than the similarity with the other names. Lamarck generally shows higher similarity with the five key words than Chambers and Wallace. This suggest that Darwin and Lamarck remained the two most prominent figures, among students of evolution, in the cultural discourse.

4.2.2 Sentiment analysis

Figure 11 (Panels A through C) displays the evolution over time of the perceptions or sentiments about the key Darwinian concepts in English books, as well as about Darwin himself. We focus on the proximity to three categories of antonyms: Unimportant vs. Important, Bad vs. Good, and Immoral vs. Moral. These dichotomies help assessing whether Darwin's concepts gained relevance and had a positive or negative connotation in the public discourse.

Although the evidence is not clear-cut, the term Evolution is, especially after 1859, perceived as more important, moral, and good, and so is Survival. Therefore, these two key concepts in Darwin's theory not only experienced an increase in use and evolution of their meaning (especially Evolution, as described in Sect. 5.1), but also were received positively. The combination of these three changes (frequency of use, semantics, and sentiments) for some of the Darwinian concepts that we consider corroborates our argument that these ideas had a novel impact on the cultural discourse. The term Darwin also shows positive reception, with spikes around the publication of *On the Origin of Species*.

5 Conclusions

To the extent that both cultural and scientific change are major drivers of long-term economic outcomes, the investigation of how these two phenomena interact with each other promises to offer a deeper understanding of their role in enhancing growth.

We focused on one of the greatest scientific breakthroughs, the theory of evolution via natural selection of Charles Darwin, and explored its impact on the public discourse. Given the undoubted importance of Darwin's theory, there is a diffused perception that it affected culture in many different ways, from changing the interpretation of nature to influencing ideas about race and equality among humans. Existing accounts, however, largely rest on

[30] Additional word similarity rankings (available from the authors) show, consistently with the evidence in Fig. 9, that Evolution became one of the most semantically similar words to terms traditionally used to describe the origin of the world by the religious doctrine, such as Creation and Genesis. Moreover, words like Progenitor, Anthropoid and Descended raised among the words with the most similar meaning to Ape.

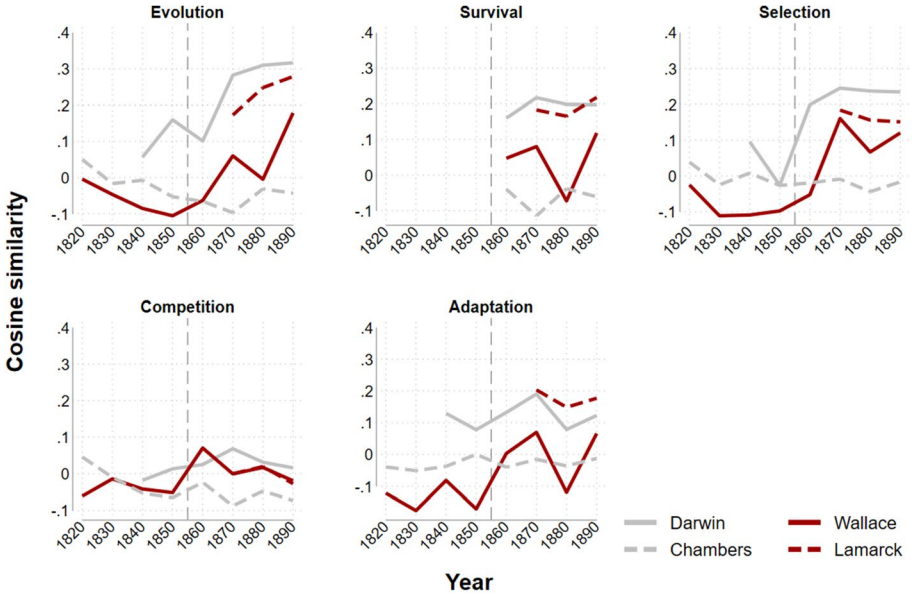
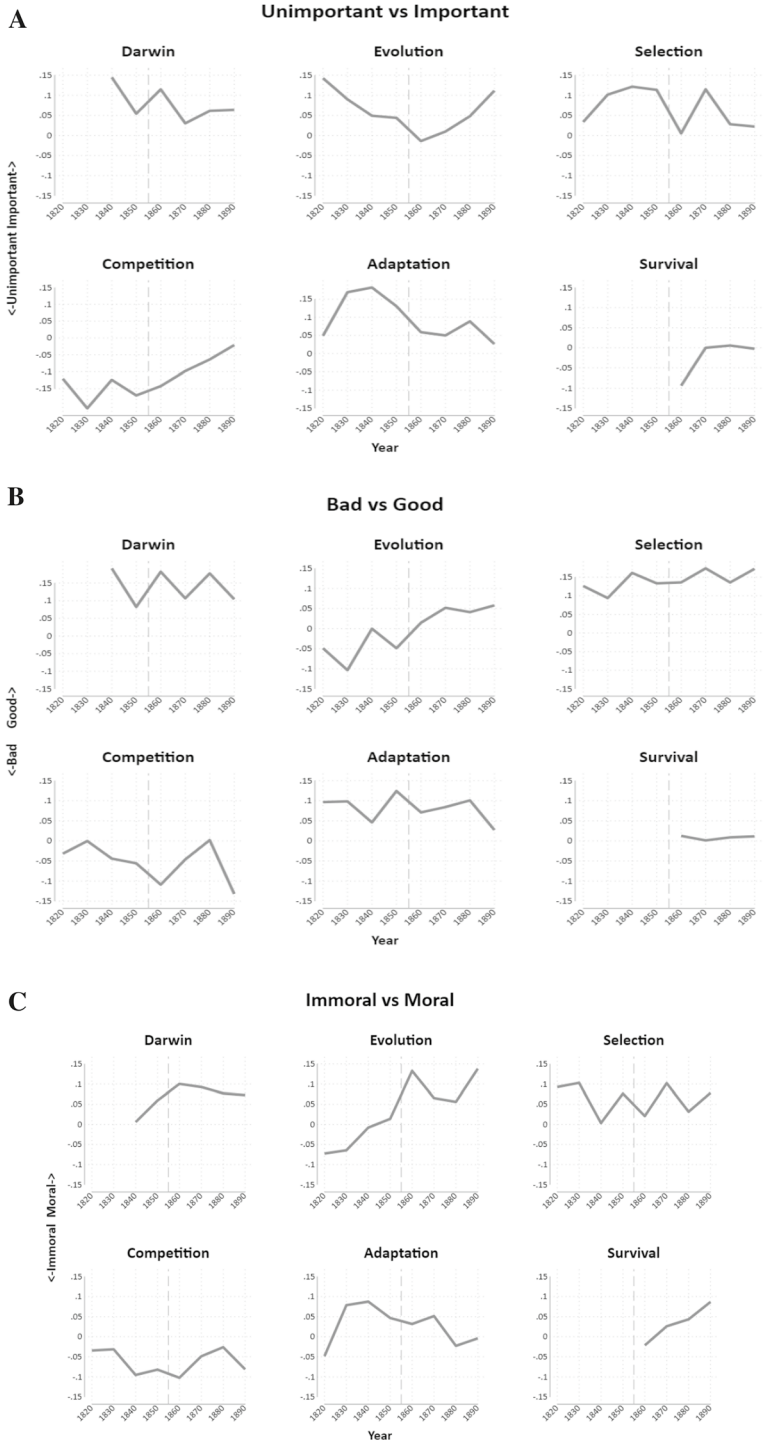


Fig. 10 Semantic Associations between the Key Words in *On the Origin of Species* and the Names Darwin, Wallace, Chambers and Lamarck. *Notes*: The graphs report the similarity between word on top of each chart and each of the four names in the legend. The weights in the word vectors were calculated with a Word2Vec algorithm. On the x-axis, 1820 represents the decade 1820–29, 1830 represents the decade 1830–39, and so on

qualitative or narrative evidence limited to scientists or cultural elites in society, whereas little is known about the wide diffusion of Darwin's ideas into society. Arguably, to affect cultural change, a scientist should have an impact on the collective imagination of a population. Moreover, it is difficult to identify, from existing accounts, which Darwinian concepts were actually novel in the cultural discourse, and which ones were already part of it. We address these challenges by analyzing the diffusion and the semantic evolution of the key words and phrases that embody Darwin's main concept in hundreds of thousands of books, with the use of techniques from machine learning. We rely on the largely unplanned publication date of *On the Origin of Species* as source of natural variation, and compare the use of these words and phrases with more generic terms that Darwin used.

Our analysis shows that the key concepts expressed by Evolution, Survival, and Natural Selection were those that diffused in fiction and non-fiction literature immediately after the publication of *On the Origin of Species*. Competition, a theme already present in the broader literature, diffused significantly more rapidly after 1859. The adoption of some of these words and phrases in the broader cultural conversation led also to a change in the meaning of the concepts, providing further evidence of the impact of Darwin's theory in society at large; overall, the attitude toward these concepts was positive rather than adversarial.

Our approach has several inductive and descriptive aspects. The choice of the concepts on which to focus may seem somewhat arbitrary; however, we based our selection on the main topics that Darwin developed, as well as on the analysis of several interpretations of Darwin's theory of evolution. Moreover, it is generally hard to provide causal identification


A
Unimportant vs Important
Darwin
Evolution
Selection
Competition
Adaptation
Survival
<-Unimportant Important->
Year
B
Bad vs Good
Darwin
Evolution
Selection
Competition
Adaptation
Survival
<-Bad Good->
Year
C
Immoral vs Moral
Darwin
Evolution
Selection
Competition
Adaptation
Survival
<-Immoral Moral->
Year

◀ **Fig. 11** Sentiment Analysis of Selected Darwinian Words in the Google Books Corpus. *Notes*: The graphs report the similarity between word on top of each chart, and set of antonyms within a certain category. On the y-axis positive values of the cosine indicate higher similarity with the "positive" end of a category (Important, Good, Moral), whereas negative values indicate closer association with the negative end (Bad, Unimportant, Immoral). On the x-axis, 1820 represents the decade 1820–29, 1830 represents the decade 1830–39, and so on

with this type of analysis. The unplanned publication date of *On the Origin of Species,* the reliance on very large amount of data, and the consistency in the patterns of different words, phrases and concepts, give us some confidence about the nature of the patterns that we established.

Finally, this is a single case study, and generalizations about the relationship between major scientific discoveries and their cultural reception are difficult to make. We limited our analysis of the impact of Darwin's theory to the diffusion on specific ideas into the broader public discourse; as such, in addition to not claiming that our work inform on how any scientific breakthrough pervades cultural attitudes, we are careful in implying that our evidence identifies an impact of Darwin's theory on culture in general. Our contribution is, in fact, to identify empirical approaches that allow for both measurement on otherwise hard-to-measure phenomena, and to propose credible strategies to assess the relationships of interest. We believe that similar approaches enabled by machine learning techniques do provide promising tools to explore this relationship beyond the specific historical episode on which we focus. Examples of relevant scientific breakthroughs include the theory of relativity or the indeterminacy principle in physics, the discovery of the DNA, and the emergence of biotechnology and genetic engineering. In fact, one could go beyond scientific discoveries and employ a similar approach to explore the cultural antecedents and effects of new technologies as well as of new industries, such as computers and the Internet (see for example Turner, 2010).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10887-022-09204-6.

Acknowledgements We thank Dora Costa, Ryan Heuser, Graeme Hirst, Xander Manshel and Yang Xu for their suggestions; and participants to presentations at Brown University, the University of Toronto, the University of Munich, the NBER Productivity Lunch, the 2018 REER Conference at Georgia Tech, the Workshop in Memory of Luigi Orsenigo at Bocconi University, the 2019 NBER Summer Institute, the 2018 Academy of Management Annual Meetings, the Virtual Economic History Seminar Series, and Monash University for their helpful feedback.

Funding We gratefully acknowledge the financial support of the National Bureau of Economic Research through the Innovation Policy Grants Program.

References

Aiden, E. and Michel, J.B. (2014). *Uncharted: Big data as a lens on human culture*. Penguin.

Alesina, A., & Giuliano, P. (2015). Culture and institutions. *Journal of Economic Literature, 53*(4), 898–944.

Armstrong, N. (1987). *Desire and domestic fiction: A political history of the novel*. Oxford University Press.

Balsmeier, B., Li, G. C., Assaf, M., Chesebro, T., Zang, G., Fierro, G., Johnson, K., Lück, S., O'Reagan, D., Yeh, B., & Fleming, L. (2018). Machine learning and natural language processing on the patent corpus: Data, tools, and new measures. *Journal of Economics & Management Strategy, 27*(3), 535–553.

Bandiera, O., Hansen, S., Prat, A., and Sadun, R. (2017). CEO Behavior and Firm Performance (No. w23248). National Bureau of Economic Research.
Bellego, C., and Pape, L. D. (2019). Dealing with the log of zero in regression models (No. 2019–13). Center for Research in Economics and Statistics.
Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems, 29*, 4349–4357.
Bush, V. (1945). *Science, the endless frontier: A report to the President*. US Govt. print.
Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186.
Cartwright, J. H. and Baker, B. (2005). Literature and science: Social impact and interaction. *Abc-Clio.*
Catalini, C., Lacetera, N., & Oettl, A. (2015). The incidence and role of negative citations in science. *Proceedings of the National Academy of Sciences, 112*(45), 13823–13826.
Chapple, J. (1986). *Science and Literature in the 19th Century*. Macmillan.
Ciccarelli, C., & Nuvolari, A. (2015). Technical change, non-tariff barriers, and the development of Italian locomotive industry, 1850–1913. *The Journal of Economic History, 75*(1), 860–888.
Cohen, M. (1999). *The sentimental education of the novel*. Princeton University Press.
D'Amico, L., & Tabellini, M. (2017). Measuring attitudes towards immigration using local newspapers' data and congressional speeches. *Working Paper*.
Davies, M. (2010). The corpus of contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing, 25*(4), 447–464.
Desmond, A. J., and Moore, J. (1994). *Darwin*. WW Norton & Company.
Dubossarsky, H., Tsvetkov, Y., Dyer, C., and Grossman, E. (2015). A bottom up approach to category mapping and meaning change. *NetWordS*, pp. 66–70.
Enke, B. (2018). Moral values and voting. *Journal of Political Economy, 128*(10), 3679–3729.
Fetter, F. W. (1975). The influence of economists in Parliament on British legislation from Ricardo to John Stuart Mill. *Journal of Political Economy, 83*(5), 1051–1064.
Fuller, R. (2017). *The book that changed America: How Darwin's theory of evolution ignited a nation*. Penguin.
Galor, O. (2011). *Unified growth theory*. Princeton University Press.
Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2017). Word embeddings quantify 100 years of gender and ethnic stereotypes. *arXiv preprint* arXiv:1711.08412.
Gentzkow, M., Kelly, B. T., and Taddy, M. (2018). Text as data. *Journal of Economic Literature*, forthcoming.
Gentzkow, M., & Shapiro, J. (2010). What drives media slant? Evidence from US daily newspapers. *Econometrica, 78*(1), 35–71.
Gerow, A., Hu, Y., Boyd-Graber, J., Blei, D. M., & Evans, J. A. (2018). Measuring discursive influence across scholarship. *Proceedings of the National Academy of Sciences, 115*(13), 3308–3313.
Gianquitto, T., & Fisher, L. (Eds.). (2014). *America's Darwin: Darwinian theory and US literary culture*. University of Georgia Press.
Giuliano, P., and Nunn N. (2021). Understanding cultural persistence and change. *Review of Economic Studies*, forthcoming
Gramsci, A. (1948). *Selections from the prison notebooks*. London: The civil society reader. University Press of New England.
Gray, A. (1860). Darwin on the origin of species. *The Atlantic*, July issue.
Guiso, L., Sapienza, P., & Zingales, L. (2004). The role of social capital in financial development. *American Economic Review, 94*(3), 526–556.
Guiso, L., Sapienza, P., & Zingales, L. (2006). Does culture affect economic outcomes? *Journal of Economic Perspectives, 20*(2), 23–48.
Gulordava, K. and Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the google books Ngram corpus. In Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics, pp 67–71. Association for Computational Linguistics.
Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016a). Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint* arXiv:1605.09096.
Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016b). Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In proceedings of the conference on empirical methods in natural language processing. Conference on empirical methods in natural language processing, 2016b, pp. 2116–2121.
Heuser, R., & Le-Khac, L. (2011). Learning to read data: Bringing out the humanistic in the digital humanities. *Victorian Studies, 54*(1), 79–86.

Heuser, R. (2016). Word vectors in the eighteenth century. *IPAM workshop: Cultural Analytics*.
Huxley, T. (1859). Darwin on the Origin of species. *The Times*, 26 December: 8–9.
Jatowt, A. and Duh, K. (2014). A framework for analyzing semantic change of words across time. In Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference, pp. 229–238. IEEE.
Jelveh, Z., Kogut, B., and Naidu, S. (2014). Detecting latent ideology in expert text: Evidence from academic papers in economics. In proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1804–1809.
Jenkins, J., Russell, W., & Suci, G. (1958). An Atlas of semantic profiles for 360 words. *American Journal of Psychology, 71*(4), 688–699.
Jones, C. I. (2002). Sources of U.S. economic growth in a world of ideas. *American Economic Review, 92*(1), 220–239.
Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis, 33*, 171–185.
Kelly, B., Papanikolaou, D., Seru, A., & Taddy, M. (2021). Measuring technological innovation over the long run. *American Economic Review: Insights, 3*(3), 303–20.
Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., and Petrov, S. (2014). Temporal analysis of language through neural language models. *arXiv preprint* arXiv:1405.3515.
Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences, 104*(12), 5241–5245.
Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review, 84*(5), 905–949.
Kulkarni, V., Al-Rfou, R., Perozzi, B., and Skiena, S. (2015). Statistically significant detection of linguistic change. In Proceedings of the 24th International Conference on World Wide Web, pp 625–635. International World Wide Web Conferences Steering Committee.
Landes, D. S. (1999). *The wealth and poverty of nations, why some are so rich and some are so poor*. WW Norton & Company.
Lansley, C. M. (2016). Charles Darwin-s debt to the Romantics. *PhD thesis, University of Winchester*.
Lévi-Strauss, C. (1963). *Structural anthropology*. Basic Books.
Levy, O., Goldberg, Y. and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Trans. ACL, 3*.
Lin, Y., Michel, J.-B., Aiden, E. L., Orwant, J., Brockman, W., and Petrov, S. (2012). Syntactic annotations for the Google books Ngram corpus. In Proceedings of the ACL 2012 system demonstrations, pp. 169–174. Association for Computational Linguistics.
Lyons, M. (2003). New readers of the nineteenth century: Women, children, workers. In G. Cavallo & R. Chartier (Eds.), *A history of reading in the West.* University of Massachusetts Press.
MacKinnon, J., & Webb, M. (2018). The wild bootstrap for few (treated) clusters. *The Journal of Econometrics, 21*(2), 114–135.
Manovich, L. (2009). *Cultural analytics: visualising cultural patterns in the era of more media*. Domus March.
Markevich, A., & Zhuravskaya, E. (2018). The economic effects of the abolition of Serfdom: Evidence from the Russian empire. *American Economic Review, 108*(4), 1075–1117.
Masci, D. (2019). *Darwin in America*. Pew Research Center.
Mayr, E. (1982). *The growth of biological thought: Diversity, evolution, and inheritance*. Harvard University Press.
Mayr, E. (1995). Darwin's impact on modern thought. *Proceedings of the American Philosophical Society, 139*(4), 317–325.
Mayr, E. (2001). The philosophical foundations of Darwinism. *Proceedings of the American Philosophical Society, 145*(4), 488–495.
McCloskey, D. N. (2016). *Bourgeois equality: How ideas, not capital or institutions, enriched the world* (Vol. 3). University of Chicago Press.
McPherson, E. G. (1942). Reporting the debates of congress.
Michalopoulos, S., & Xue, M. M. (2021). Folklore. *Quarterly Journal of Economics, 136*(4), 1993–2046.
Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*(6014), 176–182.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 3111–3119.
Mokyr, J. (2010). *The enlightened economy an economic history of Britain 1700–1850*. Yale University Press.
Mokyr, J. (2013). Cultural entrepreneurs and the origins of modern economic growth. *Scandinavian Economic History Review, 61*(1), 1–33.
Mokyr, J. (2016). *A culture of growth: The origins of the modern economy*. Princeton University Press.

Moretti, F. (2013). *Distant reading*. Verso Books.
Nguyen, D., Liakata, M., DeDeo, S., Eisenstein, J., Mimno, D., Tromble, R., & Winters, J. (2020). How we do things with words: Analyzing text as social and cultural data. *Frontiers in Artificial Intelligence, 3*, 62.
Otis, L. (2009). *Literature and science in the nineteenth century: An anthology*. Oxford University Press.
Pakes, A., & Sokoloff, K. L. (1996). Science, technology, and economic growth. *Proceedings of the National Academy of Sciences, 93*(23), 12655–12657.
Pechenick, E. A., Danforth, C. M., & Dodds, P. S. (2015). Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS One, 10*(10), e137041.
Richards, R. J. (2013). *The impact of German romanticism on biology in the nineteenth century. The impact of idealism: The legacy in philosophy and science*. Cambridge University Press.
Romer, P. M. (1990). Endogenous technological change. *Journal of political Economy, 98*(5), S71–S102.
Roodman, D., Nielsen, M. Ø., MacKinnon, J. G., & Webb, M. D. (2019). Fast and wild: Bootstrap inference in stata using boottest. *The Stata Journal, 19*(1), 4–60.
Roth, S. (2014). Fashionable functions: A Google Ngram view of trends in functional differentiation (1800–2000). *International Journal of Technology and Human Interaction, 10*(2), 35–58.
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533–536.
Schiller, R. J. (2017). Narrative economics. *American Economic Review, 107*(4), 967–1004.
Scholnick, R. (2015). *American literature and science*. University Press of Kentucky.
Sen, A. (2004). How does culture matter? In V. Rao (Ed.), *Culture and public action.* Orient Blackswan.
Spolaore, E. (2020). Commanding nature by obeying her: A review essay on Joel Mokyr's a culture of growth. *Journal of Economic Literature, 58*(3), 777–792.
Spolaore, E. (2014). *Culture and economic growth*. Edward Elgar Publishing.
Stephan, P. E. (2012). *How economics shapes science* (Vol. 1). Harvard University Press.
Thompson, B., Roberts, S. G., & Lupyan, G. (2020). Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour, 4*(10), 1029–1038.
Turner, F. (2010). *From counterculture to cyberculture: Stewart Brand, the whole earth network, and the rise of digital utopianism*. University of Chicago Press.
Whorf, B. L. (1956–2012). Language, thought, and reality: Selected writings of Benjamin Lee Whorf. MIT press.
Wilkens, M. (2015). Digital humanities and its application in the study of literature and culture. *Comparative Literature, 67*(1), 11–20.
Williamson, O. (2000). The new institutional economics: Taking Stock, looking ahead. *Journal of Economic Literature, 38*(3), 595–613.
Winans, R. B. (1975). The growth of a novel-reading public in late-eighteenth-century America. *Early American Literature, 9*(3), 267–275.
Yin, Y., Dong, Y., Wang, K., Wang, D., & Jones, B. (2021). Science as a public good: Public use and funding of science (No. w28748). National Bureau of Economic Research.