



Allotaxonomy and rank-turbulence divergence: a universal instrument for comparing complex systems

Peter Sheridan Dodds^{1,2,3*} , Joshua R. Minot¹, Michael V. Arnold¹, Thayer Alshaabi¹, Jane Lydia Adams¹, David Rushing Dewhurst¹, Tyler J. Gray^{1,4}, Morgan R. Frank⁵, Andrew J. Reagan⁶ and Christopher M. Danforth^{1,4} 

*Correspondence:

peter.dodds@uvm.edu

¹Computational Story Lab, Vermont Complex Systems Center, MassMutual Center of Excellence for Complex Systems and Data Science, Vermont Advanced Computing Center, University of Vermont, Burlington, VT 05405, USA

²Department of Computer Science, University of Vermont, Burlington, VT 05405, USA

Full list of author information is available at the end of the article

Abstract

Complex systems often comprise many kinds of components which vary over many orders of magnitude in size: Populations of cities in countries, individual and corporate wealth in economies, species abundance in ecologies, word frequency in natural language, and node degree in complex networks. Here, we introduce ‘allotaxonomy’ along with ‘rank-turbulence divergence’ (RTD), a tunable instrument for comparing any two ranked lists of components. We analytically develop our rank-based divergence in a series of steps, and then establish a rank-based allotaxonograph which pairs a map-like histogram for rank-rank pairs with an ordered list of components according to divergence contribution. We explore the performance of rank-turbulence divergence, which we view as an instrument of ‘type calculus’, for a series of distinct settings including: Language use on Twitter and in books, species abundance, baby name popularity, market capitalization, performance in sports, mortality causes, and job titles. We provide a series of supplementary flipbooks which demonstrate the tunability and storytelling power of rank-based allotaxonomy.

1 Introduction

1.1 Instruments that capture complexity

Science stands on the ability to describe and explain, and precise quantification must ultimately secure any true understanding [1]. Description itself rests on well-defined, reproducible methods of measurement, and over thousands of years, people have generated many national museums’ worth of physical and mathematical instruments along with fundamental units of measurement.

Many instruments measure a single scale. In a plane’s cockpit, barometers, altimeters, and thermometers report pressure, height, and temperature. And like a pilot flying a plane, by using human-comprehensible dashboards of single-dimension instruments, we are consequently able to successfully monitor and manage certain complex systems and processes.

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

But for complex phenomena made up of a great many types of components of greatly varying ‘size’ (clarified below in Sect. 1.2)—languages, ecologies, stock markets—we must confront two major problems with our dashboards of simple instruments [2].

First, in the face of system scale, dashboards become overwhelming. We find ourselves in high-dimensional, rapidly reconfiguring cockpits with instruments constantly appearing and disappearing. We need meters for every species, every company, every word. As a consequence, we routinely reduce a system’s description to a few summary statistics, and often to only one [3]. We quantify the massive complexity of intellect through intelligence quotients and grade point averages, health through body mass index, the complexity of civilizations by one number [4], and arguably anything by monetary value as an encoding of belief. Of course, for some systems, dimension reduction is possible and we have essential techniques for doing so such as principal component analysis [5]. Relevant to our work here, information theoretic measures such as Shannon’s entropy or the Gini coefficient are conspicuous single-number quantifications used across many fields, whether or not there is any meaningful connection to the optimal encoding of symbols for signal transmission [6–9].

Second, enabling an ability to discern change is evidently an elemental feature of any scientific instrument. Broken altimeters are a staple of stories where something goes wrong with a plane (a plane-in-trouble is the larger story trope unto itself). While tracking changes in simple measures and statistics is essential (the Dow Jones is up; today is warmer than yesterday), the cognitive trap of the single number measurement means we miss seeing the internal dynamics, and this is especially true when global statistics are constant (the Dow Jones is unchanged but stocks were volatile; today’s temperature is the same as yesterday but it’s now raining and windy).

To contend with scale and the internal diversity of complex systems, we need comprehensible, dynamically-adjusting dashboards. For comparisons of complex systems, we will argue for dynamic dashboards that have two core elements [10]:

1. A ‘big picture’ map-like overview;
- and
2. A ranking of the most ‘important’ components afforded by a tunable measure that is as straightforward as possible.

To help with our framing, we introduce a terminology family. We will use ‘allotaxonomy’ (*other order*) to mean the general comparison of the structures of two complex systems; ‘allotaxonometrics’ to refer to quantified allotaxonomy; and ‘allotaxonometers’ and ‘allotaxonographs’ for the instruments of allotaxonometrics.

1.2 Size rankings, Zipf’s law, and rank turbulence

While the instrument we develop here will have broader application, its construction focuses on two regular features of complex systems: Heavy-tailed ‘size-rank distributions’ (rather than laws), and what we will call ‘rank turbulence’—a phenomenon of system-system comparison. We describe and discuss these two signatures in turn.

We will consider systems where each component type τ has at least one measurable—and hence rankable—‘size’ s_τ . To help make clear what we mean by component types and component sizes, we list some examples in the realms of language, ecology, and stock markets.

Example types: ‘the’, ‘love’, and ‘spork’.

Type size: Number of times a word appears in a book.

In “Pride and Prejudice”, for example, the corresponding counts are 4058, 90, and 0. In linguistics, the component type and component size distinction is referred to as types and tokens [11], while more abstractly in semiotics, we have the signifier and the signified.

Example type: The species *Ornithorhynchus anatinus*, the platypus.

Type size: The number of platypuses (‘instances’ of the species) living in Australia in the wild.

Given an ecology, species may be ranked by their population numbers.

Example types: The publicly traded companies of Apple and Microsoft.

Type size: Market capitalization.

Apple and Microsoft may be viewed as components of the publicly-owned corporate world. The sizes of corporations may be broken down into many rankable dimensions such as annual revenue or number of employees worldwide. Market capitalization represents a kind of current collective belief in terms of money.

The three examples above show some of the range of what size can mean, and why we cannot readily lift terminology out from one domain to apply to all. Size for a word in a corpus means the number of indistinguishable instances of that word (many identical entities—tokens); size for species means the number of ‘biological replications’ of an individual type (many genetically similar entities of varying ages); and size for a corporation means a monetary value (one entity).

Some further examples of component size are rate (e.g., appearance of words in streaming text), physical dimensions (e.g., typical animal length or weight for a species), social popularity (e.g., number of plays of a song), scoring in sports by individual players, and so on.

We make clear that we may have no knowledge of the underlying component sizes—we may only have rankings (e.g., book rankings provided by a seller, but with numbers of sales withheld). Our core instrument functions only on rankings, incorporating size data (if known) for minor diagnostics.

When a system’s component types are ranked in descending order of some size s , we will write the size of the r th ranked component as s_r . The function s_r is commonly termed a system’s ‘Zipf distribution’ [12]. Here, we will refer to such “a component ranking by decreasing size” as a “size ranking” or, for brevity and given our paper’s framing, simply a “ranking”. (We advocate for the plainspoken naming of scientific concepts, in part to avoid misattribution [13] and with the belief that scientific truths can and should be meaningfully named for what they are.)

Though ranking is a widespread, everyday concept, the associated language can be confusing: ‘High rank’ means low r , and ‘low rank’ means high r . The highest rank size is thus s_1 . (We accommodate tied ranks per Sect. 2.1 below.)

Ranked sizes of components of complex systems commonly present (at least approximately) a decaying power law [12, 14–16]. That is, the size s_r of the r th ranked component scale as $s_r \sim r^{-\zeta}$ where $\zeta > 0$. The case $\zeta = 1$ has come to be generally referred to as Zipf’s law [12]. The corresponding frequency distribution for component sizes will behave as $f(s) \sim s^{-\gamma}$ where $\gamma = 1 + 1/\zeta > 1$.

Power laws and their discontents aside, examples of heavy-tailed size-rank distributions abound, with a few examples including word and phrase frequency in language [17, 18],

city populations [12], node degrees in scale-free networks [19], firm size [20], and numbers of dependencies for software packages [21].

We emphasize that our instrument is of use for comparing more general complex systems, for which we need only a reasonably diverse set of component types, and for which the size ranking s_r may bear any kind of heavy-tailed distribution. Below, we will explore systems with maximum component rank between roughly $10^{2.5}$ and 10^9 .

There have been two persistent criticisms of Zipf's law, one unfounded, the other true but misleading and central to our work here. The first is that Zipf's law is a meaningless artifact that arises for free through randomness [22, 23]; this is negated by a simple analysis [24], and moreover, theories of generative mechanisms have long been elaborated and tested (and contested) with the rich-get-richer mechanism proving to be a pervasive underlying algorithm [14, 21, 25, 26].

The second enduring criticism is that Zipf's exponent ζ does not vary measurably, whether it be over time for a given system or across comparable systems. Zipf's law is often plotted with an unadorned rank r on the horizontal axis, but each rank represents a component type from some vastly higher dimensional space of elements: A language's lexicon, species in an ecology, corporations in an economy.

Thus, even if two meaningfully comparable systems match exactly in a given size ranking s_r , there may well be a rich variation in the ordering of components [17, 27]. With this understanding, in earlier work by our group on comparing size rankings of word usage in large-scale texts, we introduced the concept of "lexical turbulence" [27]. We showed that in comparing word usage across decades in the Google Books English Fiction (GBEF) corpus, the flux of words across rank boundaries—rank flux ϕ_r —increased as $\phi_r \sim r^\nu$ (we found a break in scaling which we set aside here for simplicity [28, 29]). We observed superlinear scaling for rank flux with $\nu > 1.2$: Common words are relatively stable in rank, rare words much more unstable.

Here, we expand from the text-specific concept of lexical turbulence to a general one of 'rank turbulence', which in turn will help motivate our formulation of a pragmatic 'rank-turbulence divergence'.

1.3 Motivation for a rank-based divergence

In comparing complex systems, why should we use component size ranks rather than probabilities or rates? Indeed, we may select from a smorgasbord of ways to compare two probability distributions for categorical data [30–32]. Ref. [31] catalogs around 60 probability-based comparisons which are variously distances, divergences, similarities, fidelities, and inner products. And Ref. [32] details three sprawling, interrelated, single-parameter families of information-theoretic divergences.

Five main reasons push us away from probability-based divergences and towards creating and using rank-based divergences.

First, normalization problems may arise from subsampling heavy-tailed distributions [17, 33]. In natural ecological systems, for example, estimating the total number of organisms is famously difficult [33–36]. We can only then speak of relative rates and not absolute rates, and even then only for common enough species. For Twitter, for example, subsampling n -grams—phrases containing n consecutive words and/or other text elements—allows for robust estimation of the rates of common n -grams but not rare ones.

Second, not all component type characteristics can be construed (or misconstrued) as probabilities or rates. For example, rankings for many kinds of sports—at the team and player level, and not discounting the role of chance—derive from scores achieved through repeated competition [37–39].

Third, in comparison with probability-based rankings, we are able to more easily contend with components that appear in only one of two systems under comparison. We demonstrate this visualization feature as we build rank-turbulence divergence (RTD) in the following sections.

Fourth, rank orderings potentially allow for powerful and robust non-parametric statistical measures such as the standard rank correlation coefficient. All told, while in moving from sizes to rankings we may trade information for simplification, we still preserve a great deal of meaningful structure. We also expect rankings to be generally less susceptible to perturbations and errors in measurement.

Fifth and finally, rankings are an easily interpretable, ubiquitous construct, familiar to many. Ranked lists suffuse media surrounding entertainment (e.g., box office), music (Billboard charts), and sports. Indeed, we will rank anything we believe we can rank along a wide range of (often questionable) dimensions and composite scores: Individuals (wealth, fame), countries (GDP, freedom, safety, Olympic medals), cities (liveability, poverty), corporations (market capitalization, environmental records, workplace experience), universities (endowments, number of Nobel prize winners), students (grades), and animals (intelligence, dangerousness).

The above notwithstanding, distances based on comparisons of size rankings are to our knowledge relatively few, focus on traditional comparative metrics like Kendall's Tau and Spearman's rank correlation coefficient [40], and seem limited in application to extremely small systems, for example, comparing the top 20 to 50 ranked hits from two different search engines [40–42].

And while we have argued for a rank-turbulence divergence here, we nevertheless have separately constructed and explored a probability-turbulence divergence in Ref. [43]. Analogous in construction to rank-turbulence divergence, we show that probability-turbulence divergence is more sensitive to detailed system changes, has distinct limiting behavior, and corresponds to a suite of extant divergences.

1.4 Paper outline

In Sect. 2, we develop rank-turbulence divergence by (1) Establishing our notation and ranking process (Sect. 2.1); (2) Creating and explaining a specific kind of rank-rank histogram (Sect. 2.2); (3) Declaring a set of desired features for rank-turbulence divergence (Sect. 2.3); and then (4) Building and refining a rank-turbulence divergence that effectively captures these features (Sect. 2.4).

In Sect. 3, we use all of these elements to realize rank-turbulence divergence as a tunable instrument for complex system comparison through rank-turbulence divergence allotaxonographs. To both support our general explanation and explore systems in their own right, we consider comparisons at different points in time for four case studies: 1. Daily word use on Twitter, 2. Tree species abundance, 3. Baby names in the US, and 4. Market capitalization for companies.

To help demonstrate the tunability of rank-turbulence divergence and its behavior over time for dynamically evolving complex systems, we provide a suite of 'Flipbooks' of allotaxonographs as supplementary online material on the arXiv and as part of the paper's

online appendices: <http://compstorylab.org/allotaxonomy/>. Our Flipbooks expand on the paper’s allotaxonomic analyses to include season point tallies for players in the National Basketball Association (NBA); word usage in the Google Books corpus; word usage in the seven Harry Potter books; causes of death; and job advertisements. As a guide, we outline all Flipbooks in Sect. 4.

We present details of datasets and code in Sect. 5, and we round off our paper with some concluding thoughts in Sect. 6.

2 Rank-turbulence divergence

2.1 Notation, ranking methodology, and exclusive types

As mentioned in the introduction, we use simple size ranking [12], ordering a system Ω ’s types from largest to smallest size according to some measure (number, probability, mammalian fur density, etc.). Again, we write s_τ for the size of component type τ . We further indicate the rank of type τ as r_τ , and the ordered set of all types and their ranks as R_Ω .

In the case of ties, we use the conventional tied rank method of fractional ranking. For all types with the same size, we assign the mean of the sequence of ranks these types would occupy otherwise. Retaining tied information in this way makes for more sensible analytic treatment (e.g., the sum of all ranks for N types will be $\frac{1}{2}N(N + 1)$, regardless of ties). Ties (and near ties) will be important for our visualizations of rank-turbulence divergence.

Now, given two systems, Ω_1 and Ω_2 , both comprised of component types (e.g., the species of two ecosystems) of varying and rankable size (e.g., number of individuals in a species), we express rank-turbulence divergence between these systems as

$$D_\alpha^R(\Omega_1 \parallel \Omega_2). \tag{1}$$

In Sect. 2.4, we will establish α as a single tunable parameter with $0 \leq \alpha < \infty$.

Whatever complexities these systems may contain—such as networks of components—we are implicitly leaving them aside, but elaborations of our instrument will allow for their incorporation. If we have two ranked lists to compare, R_1 and R_2 , we will more directly write $D_\alpha^R(R_1 \parallel R_2)$.

The divergences we will consider here will all be expressible as linear sums of per-type contributions, meaning we can write:

$$D_\alpha^R(R_1 \parallel R_2) = \sum_{\tau \in R_{1,2;\alpha}} \delta D_{\alpha,\tau}^R(R_1 \parallel R_2). \tag{2}$$

We sort types by descending contribution (which will all be positive), $\delta D_{\alpha,\tau}^R(R_1 \parallel R_2)$, indicating this ordering by the set $R_{1,2;\alpha}$, the appropriately sequenced union of the types from both systems.

For the large-scale systems we are interested in, we expect that the overlap of types between any two systems will be partial, and generally far from complete. Hashtags on Twitter for example are constantly being invented, along with myriad lexical peculiarities (keyboard mashings, misspelling, mistypings, and more [44]).

Therefore, when comparing two systems, we extend the list of types in both systems to be the union of the types for both. If sizes are known, the sizes of types not present in a

system will be zero. We will then naturally assign the same equal last rank to all types that appear in one system and not the other.

We call types that are present in one system only ‘exclusive types.’ When warranted, we will use expressions of the form $\Omega^{(1)}$ -exclusive and $\Omega^{(2)}$ -exclusive to indicate to which system an exclusive type belongs.

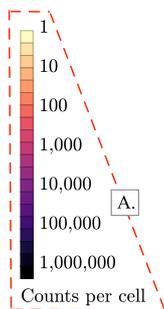
2.2 Rank-rank histograms for basic allotaxonomy

In Fig. 1, we show an example of our base system-system comparison plot, which we will call a ‘rank-rank histogram.’ We compare word usage on two days of Twitter: The day after the 2016 US presidential election, 2016/11/09, and the second day of the Charlottesville Unite the Right rally, 2017/08/13 (see Sect. 5.1 for description of datasets). As we describe below, our histograms fully present the meaningful differences between two size-rank distributions, allowing for divergence measures to be overlaid in easily interpretable ways (c.f., [47, 48]). As an aid for the reader, we include dash-lined-bordered elements from Fig. 1 throughout our discussion.

To construct Fig. 1, we first parse tweets into 1-grams (preserving case), find 1-gram frequencies for each day, and then determine each day’s separate ranked list of 1-grams according to those frequencies. For both days, and purely by choice, we take the subset of 1-grams that contain simple latin characters. We next generate a merged list of latin character 1-grams observed on both days and thereby obtain rank-rank pairs for all 1-grams. As described above, the separately ranked lists for each day will be extended by exclusive 1-grams (i.e., those that appear on only one day). All exclusive 1-grams will be tied for the last rank on the day they do not appear.

In general, we will denote the rank of type τ in system $\Omega^{(1)}$ by $r_{\tau,1}$, and the same type’s rank in system $\Omega^{(2)}$ by $r_{\tau,2}$.

For our histograms, we bin rank-rank pairs $(r_{\tau,1}, r_{\tau,2})$ into cells uniformly in logarithmic space. Cell width is adjustable; here we choose 1/15 of an order of magnitude. We use a perceptually uniform colormap (magma [46]), with the number of rank-rank pairs per cell increasing per the lower left scale (Fig. 1A). That the rank-rank pair counts per cell reach up towards 10^6 should make clear that some form of histogram is necessary for attempting to visualize the kind of rank turbulence we see here for Twitter. A simple plot of all $(r_{\tau,1}, r_{\tau,2})$ points produces an incomprehensible density.



We orient our histograms in a diamond format, rotating the standard horizontal-vertical axes $\pi/4$ counterclockwise. We do so to eliminate a perceptual bias towards interpreting causality (separately suggested in [49]). The vertical and horizontal coordinates in the rotated histogram are proportional to $\log_{10} r_{\tau,1}r_{\tau,2}$ (measured downwards) and $\log_{10} r_{\tau,2}/r_{\tau,1}$

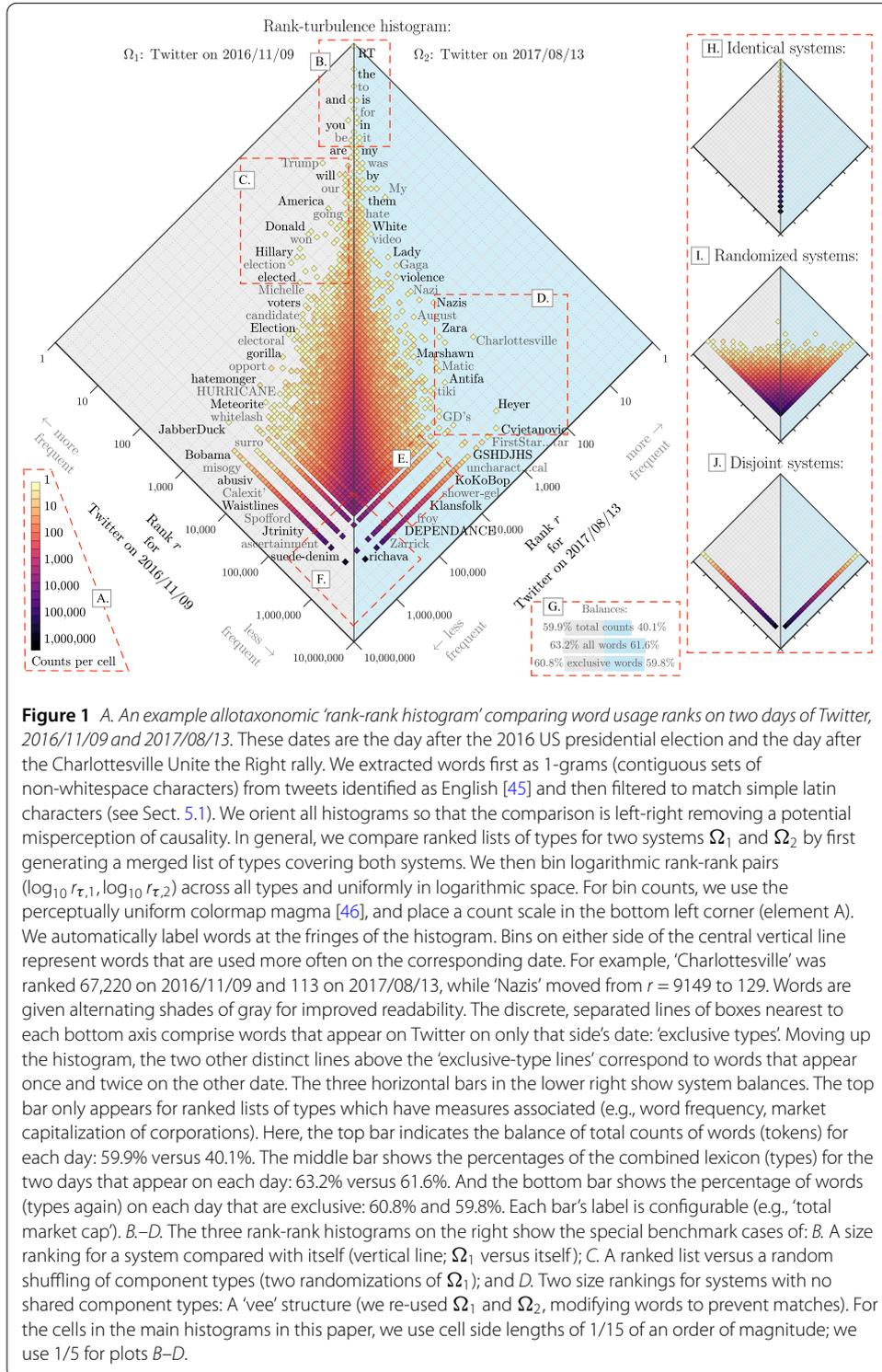


Figure 1 A. An example allotaxonomic ‘rank-rank histogram’ comparing word usage ranks on two days of Twitter, 2016/11/09 and 2017/08/13. These dates are the day after the 2016 US presidential election and the day after the Charlottesville Unite the Right rally. We extracted words first as 1-grams (contiguous sets of non-whitespace characters) from tweets identified as English [45] and then filtered to match simple latin characters (see Sect. 5.1). We orient all histograms so that the comparison is left-right removing a potential misperception of causality. In general, we compare ranked lists of types for two systems Ω_1 and Ω_2 by first generating a merged list of types covering both systems. We then bin logarithmic rank-rank pairs ($\log_{10} r_{\tau,1}, \log_{10} r_{\tau,2}$) across all types and uniformly in logarithmic space. For bin counts, we use the perceptually uniform colormap magma [46], and place a count scale in the bottom left corner (element A). We automatically label words at the fringes of the histogram. Bins on either side of the central vertical line represent words that are used more often on the corresponding date. For example, ‘Charlottesville’ was ranked 67,220 on 2016/11/09 and 113 on 2017/08/13, while ‘Nazis’ moved from $r = 9149$ to 129. Words are given alternating shades of gray for improved readability. The discrete, separated lines of boxes nearest to each bottom axis comprise words that appear on Twitter on only that side’s date: ‘exclusive types’. Moving up the histogram, the two other distinct lines above the ‘exclusive-type lines’ correspond to words that appear once and twice on the other date. The three horizontal bars in the lower right show system balances. The top bar only appears for ranked lists of types which have measures associated (e.g., word frequency, market capitalization of corporations). Here, the top bar indicates the balance of total counts of words (tokens) for each day: 59.9% versus 40.1%. The middle bar shows the percentages of the combined lexicon (types) for the two days that appear on each day: 63.2% versus 61.6%. And the bottom bar shows the percentage of words (types again) on each day that are exclusive: 60.8% and 59.8%. Each bar’s label is configurable (e.g., ‘total market cap’). B–D. The three rank-rank histograms on the right show the special benchmark cases of: B. A size ranking for a system compared with itself (vertical line; Ω_1 versus itself); C. A ranked list versus a random shuffling of component types (two randomizations of Ω_1); and D. Two size rankings for systems with no shared component types: A ‘vee’ structure (we re-used Ω_1 and Ω_2 , modifying words to prevent matches). For the cells in the main histograms in this paper, we use cell side lengths of 1/15 of an order of magnitude; we use 1/5 for plots B–D.

(measured rightwards), and these are dimensions we will encounter later in our construction of rank-turbulence divergence.

Types that have higher rank in system Ω_1 will be represented by points on the left of the vertical $r_{\tau,1} = r_{\tau,2}$ line, while those with higher rank in system Ω_2 will appear on the right

side. Types falling along or near the center vertical line have the same or similar ranks in both systems.

For all rank-rank histograms we show in the main paper, we compare systems at different time points. Time moving from left-to-right is a natural choice, and will govern our arrangement of dynamically evolving systems. In general however, comparisons between two systems may not involve any left-right ordering, and the choice will be arbitrary, (e.g., comparison of word usage in two books published in the same year or species abundance in two distinct ecological systems).

We automatically annotate words along the edges of the histogram. To do so, we first specify a fixed bin size moving down the vertical axis. For each bin and each side of the plot, we find the word furthest away horizontally from the center line, i.e., the word maximizing $|\log_{10} r_{\tau,1}/r_{\tau,2}|$. Annotated words are oriented to the far side of the point $(r_{\tau,1}, r_{\tau,2})$ relative to the center, but are vertically centered by bin for overall clarity (meaning that their vertical position relative to $(r_{\tau,2}, r_{\tau,1})$ will fluctuate). For these bare histograms with no divergence measure, we also assign type names with alternating shades of gray for readability. Where more than one word is equally far away from the center, we randomly choose one as a representative example.

To aid a user’s perception of what meaning might be conferred by a rank-rank histogram, we highlight a selection of the annotated words in Fig. 1. Broadly, there are four main regions: 1. The top of the diamond; 2. The sides of the histogram; 3. The lower linear and point structures of the histogram; and 4. The bottom of the diamond.

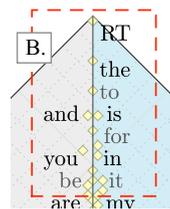


Fig. 1B: Types appearing towards the top of the diamond rank high for both systems. For Fig. 1, the 1-gram ‘RT’ is the most common word on both days: $r_{RT,1} = r_{RT,2} = 1$. Signifying retweet, ‘RT’ is an important—if Twitter-specific—functional structure, indicating the strength of social amplification on Twitter. The words ‘the’ and ‘to’ are ranked 2nd and 3rd on both dates, while ‘and’ and ‘is’ are ranked 4th and 5th on 2016/11/09 and reversed to 5th and 4th on 2017/08/13, leading to their offset locations. Such changes of high rank types will be important in analyzing many kinds of systems, and we will see later that they are only picked up by certain divergences.

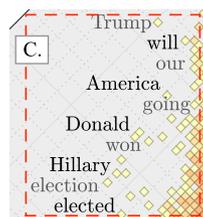


Fig. 1C: Moving down the histogram, we see that turbulence starts to become noticeable around $r = 10^2$, and we see increasingly less common and differentiating words appear. Types appearing furthest horizontally from the center vertical axis show the most rela-

tive change in rank. On 2016/11/09, ‘Trump’ stands out relative to nearby words. Further down, ‘America,’ ‘Donald,’ ‘voters,’ and ‘election’ are all clearly off-axis. On 2017/08/13, the words ‘Charlottesville’ and ‘Heyer’ are most prominent (Heather Heyer was a protester who was murdered by vehicular homicide on August 12, 2017).



Fig. 1D: While election and Charlottesville terms dominate the sides of the histogram, (seemingly) unrelated names and events also appear. On the left of the histogram and/or list, we see ‘gorilla’ and ‘Meteorite’. Harambe was a gorilla who was killed in a Cincinnati zoo after a boy entered his enclosure in 2016/05. Harambe became part of various internet memes including ones putting him forward as a write-in candidate for US president. On the right of the histogram, we find Lady Gaga and Zara Larsson (both performed concerts), and the K-pop (Korean pop) band BTS [50] which was enjoying its rise to ultrafame over this time period [51].

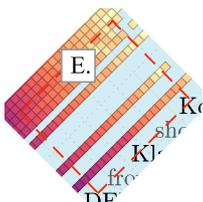


Fig. 1E: The separated lines and points at the bottom of the histogram arise from logarithmic spacing. For systems with heavy-tailed rankings of discrete-sized components, we often observe many types of the least size. Here, where type size is word count, we have many hapax legomena—words that appear only once in a corpus. For books approximately obeying Zipf’s law, the fraction of a lexicon that appears is around 1/2 [14]—the rare are legion.

Moving upwards from the bottom, the three separated lines in Fig. 1’s histogram correspond to words appearing zero times (‘exclusive types’ by our definition), once, and twice on the other side’s day.

For example, at the extreme of the lowest line on the right, we see ‘Cvjetanovic’, a $\Omega^{(2)}$ -exclusive word that is highly ranked on 2017/08/13 ($r_{Cvjetanovic,2} = 672$). The word is the last name of a member of Identity Evropa who was part of the Unite the Right Rally. A photo of Cvjetanovic holding a tiki torch and yelling was widely circulated [52]. The word ‘Cvjetanovic’ did not appear on 2016/11/09 and with zero counts, is tied with many other words that only appear on 2017/08/13 ($r_{Cvjetanovic,1} = 1,552,865$). As another example, the word ‘Heyer’ appeared once on 2016/11/09 and is consequently part of the second discrete line on the right side.

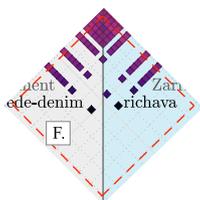


Fig. 1F: The least important and least differentiating types appear at the bottom of the histogram. These types are low rank in both systems. The bottommost annotations of Fig. 1—‘suede-denim’ and ‘richava’—appear once on the dates of their respective sides. These creatures of the lexical abyss are just two examples of on the order of 10^6 words appearing once on only one of the two dates (per the count scale, Fig. 1A).

We emphasize that types annotated at or near the bottom of the diamond cannot be important individually—no divergence measure should present ‘richava’ as a meaningful word in itself for these two days of Twitter. Even so, indicating a few examples of these rare and unimportant words along the bottom of the histogram provides a helpful check that this is indeed the case. With the aim of improving the instrument’s affordance of understanding, when we introduce rank-turbulence divergence, we will fade annotations according to type-level divergence contributions. Annotations for doubly rare types will always be strongly backgrounded.

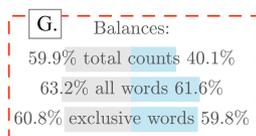


Fig. 1G: For all allotaxonographs, we show balances at the bottom right of the rank-rank histogram. Two kinds of exclusive type comparisons for the numbers of types in each system are recorded in the bottom two bars, while the top bar conveys information about type sizes. The top bar is rendered only when sizes are known (this is the only part of the allotaxonographs that is not determined purely from type rankings). In the present paper, we work from datasets with known type sizes, and all allotaxonographs will have all three bars. All balances show normalized quantities rather than absolute numbers

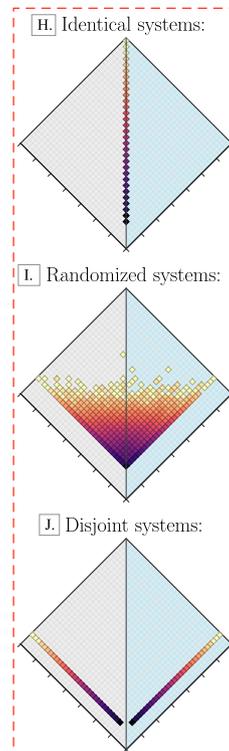
As we will see, these three balances can vary greatly across system comparisons.

The top balance bar shows the relative balance of the two systems’ sizes, if known. For our Twitter example, this top bar shows the breakdown of total counts of 1-grams (type size) between the two dates at 59.9% and 40.1%. We thus see that the election generated about 50% more 1-grams (which tracks with tweets) than events of Charlottesville.

The middle balance bar shows the fraction of types in each system as a percentage of the union of types from both systems. For Twitter, we have that of all words in the combined lexicon for the two days combined, just over 60% appear on each of the two days (63.2% and 61.6%).

The bottom balance bar shows that given a system’s set of types, what percentage of those types appear only in that system—exclusive types. For the Twitter example, we take the separate lexicons for each day, and find that around 60% of words are exclusive for both days, further giving a sense of strong turnover (60.8% and 59.8%).

We add that the script for generating allotaxonographs (`figallotaxonometer9000.m` in Matlab provided in the paper's Gitlab repository) returns some diagnostics including the underlying numbers used to compute the above balances.



Figs. 1H, I, and J show examples of three extremes of how systems might compare on rank-rank histograms.

In Fig. 1H, we compare the size ranking for identical systems (Ω_1 from Fig. 1). The outcome is a colormap version of the system's rankings binned logarithmically and arranged on the vertical $r_{\tau,1} = r_{\tau,2}$ line.

In Fig. 1I, we present the visualization of a system compared with a randomized version of itself. The nature of logarithms means that the lower triangle is well filled with density growing with increasing rank. Using a linear scale, we would see a statistically uniform histogram.

Finally, in Fig. 1J, we compare size-rank distributions for systems with completely distinct sets of types. After merging types across systems, ranking of types for each system places all types of the other system in a tie for last place. The result is two marginal size-rank distributions forming a 'vee'. We have already seen examples of these linear features in Fig. 1. If system component lists are sufficiently truncated—whether by measurement limitations or by choice—we will also see these kinds of marginal structures appear but in an inconsistent fashion. We will discuss truncation effects further in Sect. 3.6, after introducing rank-turbulence divergence.

2.3 Desirable allotaxonomic features for rank-turbulence divergence

On their own, our annotated rank-rank histograms give a map-like overview of how two systems differ. For Twitter, Fig. 1 presents a clear texture of words associated with the 2016 US election on the left and the 2017 events of Charlottesville on the right. But which words

are most important? How do we compare the relatively rare ‘Heyer’ with the common ‘My’, both words that have higher ranks on 2017/08/13?

Our goal now is to construct a rank-based divergence for comparing complex systems, one that will function as an instrument overlaying rank-rank histograms. We would like our divergence to be able to bear the following 11 descriptors, which range from concrete and simple to qualitative:

1. Rank-based: Directly built for comparing ranked lists generated by any meaningful ordering.
2. Symmetric: $D_{\alpha}^R(R_1 \| R_2) = D_{\alpha}^R(R_2 \| R_1)$.
3. Semi-positive: $D_{\alpha}^R(R_1 \| R_2) \geq 0$, and $D_{\alpha}^R(R_1 \| R_2) = 0$ only if the systems are formed by the same components with matching rankings, $R_1 = R_2$.
4. Metric-capable: Given the preceding two conditions are met, we would want D_{α}^R to also satisfy the triangle inequality.
5. Scale and unit invariant: This is automatic because rankings will not change if either one or both systems are rescaled in their entirety, or remeasured according to a different system of units.
6. Linearly separable, for interpretability. As framed in Eq. (2), each type τ additively contributes to rank-turbulence divergence a quantity $\delta D_{\alpha, \tau}^R(R_1 \| R_2)$, allowing for simple ranking of types to assess importance.
7. Subsystem applicable: Ranked lists of any principled subset may be equally well compared (e.g., hashtags on Twitter, stock prices of a certain sector, etc.).
8. Effective across system sizes, possibly size independent: While not being explicitly interpretable as certain probability divergences (e.g., Kullback-Leibler divergence), rank-turbulence divergence $D_{\alpha}^R(\Omega_1 \| \Omega_2)$ should be normalizable to allow for sensible comparisons of rank-turbulence divergences across system sizes. Linear separability means that whatever normalization we use, the ordering of contributions of individual types will be unchanged.
9. Heavy-tailed distributions: Rank-turbulence divergence should be applicable to systems with rank-ordered component size distributions that are heavy-tailed.
10. Tunable: The acknowledgment that while many stand-alone divergences exist for probability distributions [31, 32], in practice there are families of divergences on offer, and these have the potential to be adaptive and provide much more power and insight [32].
11. Storyfinding: Features 1–10 will ideally combine to help us rapidly see which types are most important in distinguishing two ranked lists.

2.4 Development of rank-turbulence divergence

With these features in mind, we move now to properly constructing our conception of rank-turbulence divergence. We begin with the observation that by definition, a type τ 's size rank is inversely related to its size. We thus will want to deal with inverses of ranks.

Given element τ has a size rank $r_{\tau,1}$ in system 1 and $r_{\tau,2}$ in system 2, a raw starting point for an element-level divergence incorporating rank inverses would be:

$$\left| \frac{1}{r_{\tau,1}} - \frac{1}{r_{\tau,2}} \right|. \tag{3}$$

As we will demonstrate later, experimentation with this fixed form reveals a bias towards types with high ranks (again, the highest rank is $r = 1$).

We modify the above expression by introducing a parameter $\alpha \geq 0$:

$$\left| \frac{1}{[r_{\tau,1}]^\alpha} - \frac{1}{[r_{\tau,2}]^\alpha} \right|^{1/\alpha}. \tag{4}$$

We now have tunability: As $\alpha \rightarrow 0$, high ranked types are increasingly dampened relative to low ranked ones. For words in texts, for example, the weight of common words and rare words will become increasingly closer together. (Our construction and its behavior are in parts resemblant of but distinct from that of generalized entropy [53–55] and Hill numbers in ecology [8, 34].)

At the other end of the dial, $\alpha \rightarrow \infty$, high rank types will dominate. For texts, function words will prevail while the contributions of rare words will vanish.

The $\alpha \rightarrow \infty$ limit will prove to be a natural parameter endpoint for rank-turbulence divergence when we realize it as an instrument, and is something we wish to preserve as we address the $\alpha \rightarrow 0$ limit.

However, the limit of $\alpha \rightarrow 0$ in Eq. (4) does not yet behave as we might hope. We see that if $r_{\tau,1} \neq r_{\tau,2}$, Eq. (4) tends towards

$$\alpha^{1/\alpha} \left| \ln \frac{r_{\tau,1}}{r_{\tau,2}} \right|^{1/\alpha}, \tag{5}$$

which in turn will tend toward ∞ as $\alpha \rightarrow 0$.

In considering how to remedy this problematic limit, we observe that Eq. (5) contains a readily interpretable structure which we have already encountered in the preceding section: The log-ratio of ranks. In Sect. 2.2, we established a graphical interpretation for the rank-rank histogram in Fig. 1. We identify $|\ln \frac{r_{\tau,1}}{r_{\tau,2}}| = |\ln r_{\tau,1} - \ln r_{\tau,2}|$ as being proportional to the horizontal distance from the $(\log_{10} r_{\tau,1}, \log_{10} r_{\tau,2})$ point to the histogram’s vertical midline.

In order to fashion a well-behaved $\alpha \rightarrow 0$ limit, while (1) preserving the core of Eq. (4), $|\frac{1}{[r_{\tau,1}]^\alpha} - \frac{1}{[r_{\tau,2}]^\alpha}|^{1/\alpha}$, (2) maintaining the form of the large α limit, and (3) only using modifications that are monotonic in α , we introduce a prefactor and adjust the exponent in Eq. (4) as follows:

$$\frac{\alpha + 1}{\alpha} \left| \frac{1}{[r_{\tau,1}]^\alpha} - \frac{1}{[r_{\tau,2}]^\alpha} \right|^{1/(\alpha+1)}. \tag{6}$$

The $\alpha \rightarrow 0$ limit is now simply $|\ln \frac{r_{\tau,1}}{r_{\tau,2}}|$, while the $\alpha \rightarrow \infty$ limit is unchanged. (We note that an alternate modification of introducing a prefactor of $\alpha^{-1/\alpha}$ to Eq. (4) fails the requirement of monotonicity.)

Finally, in summing over all types and incorporating a normalization prefactor $\mathcal{N}_{1,2;\alpha}$, we have our prototype, single-parameter rank-turbulence divergence:

$$\begin{aligned} D_\alpha^R(R_1 \| R_2) &= \sum_{\tau \in R_{1,2;\alpha}} \delta D_{\alpha,\tau}^R(R_1 \| R_2) \\ &= \frac{1}{\mathcal{N}_{1,2;\alpha}} \frac{\alpha + 1}{\alpha} \sum_{\tau \in R_{1,2;\alpha}} \left| \frac{1}{[r_{\tau,1}]^\alpha} - \frac{1}{[r_{\tau,2}]^\alpha} \right|^{1/(\alpha+1)}. \end{aligned} \tag{7}$$

Deducing the form of the normalization factor $\mathcal{N}_{1,2;\alpha}$ requires a combined analytic and numerical approach. We compute $\mathcal{N}_{1,2;\alpha}$ by taking the two systems to be disjoint while maintaining their underlying ranked lists. Thus, we ensure $0 \leq D_\alpha^R(R_1 \| R_2) \leq 1$ where the limits of 0 and 1 correspond, respectively, to the two systems having identical and disjoint size-rank distributions.

To determine $\mathcal{N}_{1,2;\alpha}$, we observe that if the size-rank distributions are disjoint, then in $\Omega^{(1)}$'s merged ranking, the rank of all $\Omega^{(2)}$ types will be $r = N_1 + \frac{1}{2}N_2$, where N_1 and N_2 are the number of distinct types in each system. Similarly, $\Omega^{(2)}$'s merged ranking will have all of $\Omega^{(1)}$'s types in last place with rank $r = N_2 + \frac{1}{2}N_1$. The normalization factor is then:

$$\begin{aligned} \mathcal{N}_{1,2;\alpha} &= \frac{\alpha + 1}{\alpha} \sum_{\tau \in R_1} \left| \frac{1}{[r_{\tau,1}]^\alpha} - \frac{1}{[N_1 + \frac{1}{2}N_2]^\alpha} \right|^{1/(\alpha+1)} \\ &\quad + \frac{\alpha + 1}{\alpha} \sum_{\tau \in R_2} \left| \frac{1}{[N_2 + \frac{1}{2}N_1]^\alpha} - \frac{1}{[r_{\tau,2}]^\alpha} \right|^{1/(\alpha+1)}. \end{aligned} \tag{8}$$

2.5 Tunability of rank-turbulence divergence: limits

We will use rank-turbulence divergence's tunability to accentuate more rare ($\alpha \rightarrow 0$) or more common types ($\alpha \rightarrow \infty$). For reference, we lay out the full expressions for these two limits, and will later see their graphical realizations. Per our construction of Eq. (7), in the limit of $\alpha \rightarrow 0$, we have

$$D_0^R(R_1 \| R_2) = \sum_{\tau \in R_{1,2;\alpha}} \delta D_{0,\tau}^R = \frac{1}{\mathcal{N}_{1,2;0}} \sum_{\tau \in R_{1,2;\alpha}} \left| \ln \frac{r_{\tau,1}}{r_{\tau,2}} \right|, \tag{9}$$

where

$$\mathcal{N}_{1,2;0} = \sum_{\tau \in R_1} \left| \ln \frac{r_{\tau,1}}{N_1 + \frac{1}{2}N_2} \right| + \sum_{\tau \in R_2} \left| \ln \frac{r_{\tau,2}}{\frac{1}{2}N_1 + N_2} \right|. \tag{10}$$

Types experiencing the largest relative change in rank will feature most strongly, and these are types that are rare in one system, and extremely common in the other. Because of the term $\ln \frac{r_{\tau,1}}{r_{\tau,2}}$, the $\alpha = 0$ limit for rank-turbulence divergence is most resemblant of the Kullback–Leibler and Jeffrey divergences [30].

In the limit of $\alpha \rightarrow \infty$, we have instead

$$\begin{aligned} D_\infty^R(R_1 \| R_2) &= \sum_{\tau \in R_{1,2;\alpha}} \delta D_{\infty,\tau}^R \\ &= \frac{1}{\mathcal{N}_{1,2;\infty}} \sum_{\tau \in R_{1,2;\alpha}} (1 - \delta_{r_{\tau,1}r_{\tau,2}}) \max_\tau \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\}. \end{aligned} \tag{11}$$

Having the lowest values of $1/r$, highest-rank types will dominate the $\alpha \rightarrow \infty$ limit. The normalization factor for $\alpha = \infty$ is:

$$\mathcal{N}_{1,2;\infty} = \sum_{\tau \in R_1} \frac{1}{r_{\tau,1}} + \sum_{\tau \in R_2} \frac{1}{r_{\tau,2}}. \tag{12}$$

First, we build upon the histogram of Fig. 1.

We use rank-turbulence divergence with $\alpha = 1/3$, as indicated on the scale in the top left of the graph. We discuss the choice of α below.

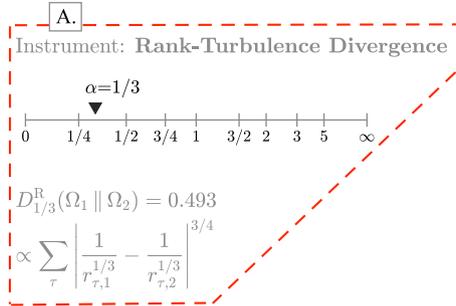


Fig. 2A: In all our divergence graphs, we include the divergence’s expression above the top left of the histogram. We display the value of the divergence, which for our Twitter example is $D_{1/3}^R(R_1 \parallel R_2) = 0.493$. We also show the core form of RTD, excluding constants of proportionality. (Our figure-making code presents formats for other kinds of divergences such as generalized entropy and probability-turbulence divergence [43].) For our own implementation of rank-turbulence divergence, we have chosen to make the increments of α discrete as multiples of $1/12$. This discretization is particularly useful for $\alpha \leq 3/2$, the range of α for which most of the variation in rank-turbulence divergence takes place. The α scale in Fig. 2A uses an inverse tangent transformation that is effective for functional use of the instrument. As we will see, near $\alpha = 0$, the list’s variation with steps of $1/12$ is not abrupt.

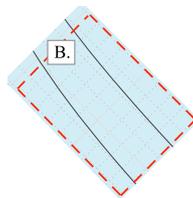
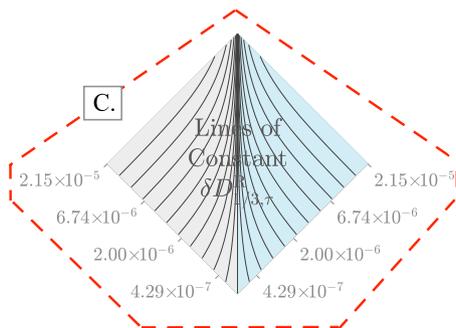


Fig. 2B: We overlay the histogram with contour lines of constant $\delta D_{1/3,\tau}^R$. The contour lines are chosen so that they are anchored at evenly spaced points along the bottom two axes, making for simple tracking as α is varied.

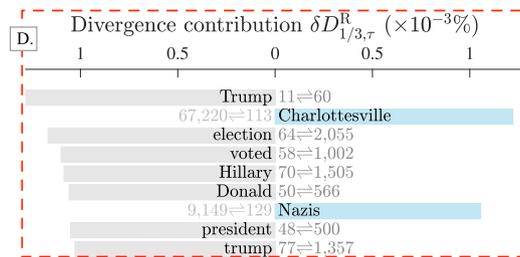
Fig. 2C: The inset to the upper right of the histogram provides a scale for values of $\delta D_{1/3,\tau}^R$, per the tick marks. This inset also shows the contour lines of the chosen instrument, matching those of the main histogram.



Our last enhancement is to foreground annotations for types based on how much they contribute to the divergence. The annotations and their locations on the histogram largely remain unchanged from Fig. 1 (some may vary because of chance in the automatic annotation). We now incorporate a linear gray scale based on $\delta D_{1/3,\tau}^R$, with higher scoring words accentuated, lower scoring words faded. We now see ‘Trump’ and ‘Charlottesville’ stand out among the histogram annotations of Fig. 2. Common words that have not changed rank (‘RT’, ‘the’, and ‘to’) as well as words rare on one day and absent on the other (‘suededenim’ and ‘richava’) have all been strongly backgrounded.

Second, we locate a list of words on the right of the instrument.

Fig. 2D: We order the top 40 words by decreasing value of $\delta D_{1/3,\tau}^R$, as indicated by the underlying bars. We orient words to the left and right in accordance with the day of their higher rank; the bar colors of light gray and light blue match the histogram’s format. Opposite each bar, we show the word’s rank on each day.



For example, we see ‘Trump’ has the highest divergence contribution overall, moving from $r = 11$ to 60. These ranks indicate a maintenance of extraordinary levels of lexical ultrafame [51]), but the drop from $r = 11$ to 60 registers more strongly for $\delta D_{1/3,\tau}^R$ than all other rank shifts. On the opposing date, ‘Charlottesville’ scores comparably to ‘Trump’ and is second overall. In contrast to ‘Trump’, however, ‘Charlottesville’ is a word that changes rank dramatically across the two dates, moving from $r = 67,220$ to 113.



Fig. 2E: It is useful to be able to see which ‘important’ (i.e., high $\delta D_{\alpha,\tau}^R$) elements are part of only one system (i.e., important exclusive types). In the ordered list, we indicate exclusive types by a directed open triangle, that will either precede a word appearing on the left or trail a word appearing on the right.

For Fig. 2 with α set at 1/3, there is only one such word in the top 40 divergence contributions: ‘Cvjetanovic’ (discussed in Sect. 2.2). For general systems, as we tune α towards zero, more single-system types will move up the list, and conversely fall back down if we instead dial α towards ∞ .



Fig. 2F: At the bottom of the word list, we indicate the percentage contribution to the divergence score from each system. Generally, we find these contributions to be close to equal.

3.2 Tuning rank-turbulence divergence allotaxonographs

For Fig. 2, we have chosen $\alpha = 1/3$ because it delivers a reasonably balanced list of words with ranks from across the common-to-rare spectrum. Our choice here is based purely on a visual inspection. We have considered several automated methods for determining an optimal α , but leave these for future work.

To demonstrate how tuning α controls the contour lines and alters the word list on a rank-turbulence divergence graph, we provide Flipbook S1 where we sweep through a set of 11 values of α in steps: $0, \frac{1}{12}, \frac{2}{12}, \frac{3}{12}, \frac{4}{12}, \frac{5}{12}, \frac{6}{12}, \frac{8}{12}, 1, 2, 5$, and ∞ . As we increase α , the set of words (and in general, types) with highest $\delta D_{\alpha, \tau}^R$ transform from being dominated by rare words to function words. Even so, a few words maintain prominence across a wide range of α . For example, ‘Trump’ is the top word for $\alpha = 1/3$ to $5/4$, dropping only to 5th for $\alpha = \infty$. (Because of its function-word-like fame, for $\alpha \leq 1/6$, ‘Trump’ does not register in the top 40.) For $0 \leq \alpha \leq 5/6$, Charlottesville-related words lead the right side of the list (‘Cvjetanovic’, ‘Heyer’, and ‘Charlottesville’). At the limit of $\alpha = \infty$, the only top 40 Charlottesville word is ‘white’ (per the prevalence of ‘white supremacists’ and similar terms).

To further our investigation, We provide two more Flipbooks for Twitter. Flipbook S2 shows how the allotaxonograph of Fig. 2 changes if we control the percentage of retweets included in our sample. In varying from 1% to 100%, we see that the texture of the election side does not change greatly—the amplified and unamplified versions of Twitter match well. However, the Charlottesville date shows that the 1% retweet sample is much more pop culture focused. As we move through Flipbook S2 and dial up to fully include all retweets for 2017/08/13, we see words surrounding the events in Charlottesville rise up the list of dominant contributions.

In Flipbook S3, we start with 2019/01/03 and compare forwards in time, roughly doubling the number of days for each step, ending with 2020/01/04, the date of the assassination of the Iranian general Soleimani by the United States. We see the topics of anchor date 2019/01/03 become more clear as the date moves further into the past: Government shutdown, the border wall, and Congresswoman Rashida Tlaib. The comparison future date travels through a wide range of events. We observe that rank-turbulence divergence slowly increases as we compare days increasingly further apart. Visually, we see the rank-rank histogram broaden subtly. Determining how an optimal α changes with time scales would be a natural part of possible future work.

To explore in more depth the value of having a tunable allotaxonomic instrument, we move away from news and Twitter to consider distributions presented by two different kinds of systems, one ecological, the other cultural: Tree species abundances and popularity of baby names.

3.3 Species abundance: example rank-turbulence divergence allotaxonograph for the limit of $\alpha = 0$

In Fig. 3, we show a rank-turbulence divergence graph comparing tropical tree species numbers on Barro Colorado Island (BCI) in the Panama Canal [61] for five-year censuses completed in 1985 and 2015 ($\Omega^{(1)}$ and $\Omega^{(2)}$) [56].

In being visually close to the limit of comparing two identical rankings ($D_0^R(R_1 \| R_2) = 0$, Fig. 1H), the histogram’s vertical linear form immediately shows that the species abundance distributions are strongly aligned. Because of the possibility of exogenous catastrophic events such as fires and the abrupt transitions accessible by complex dynamical

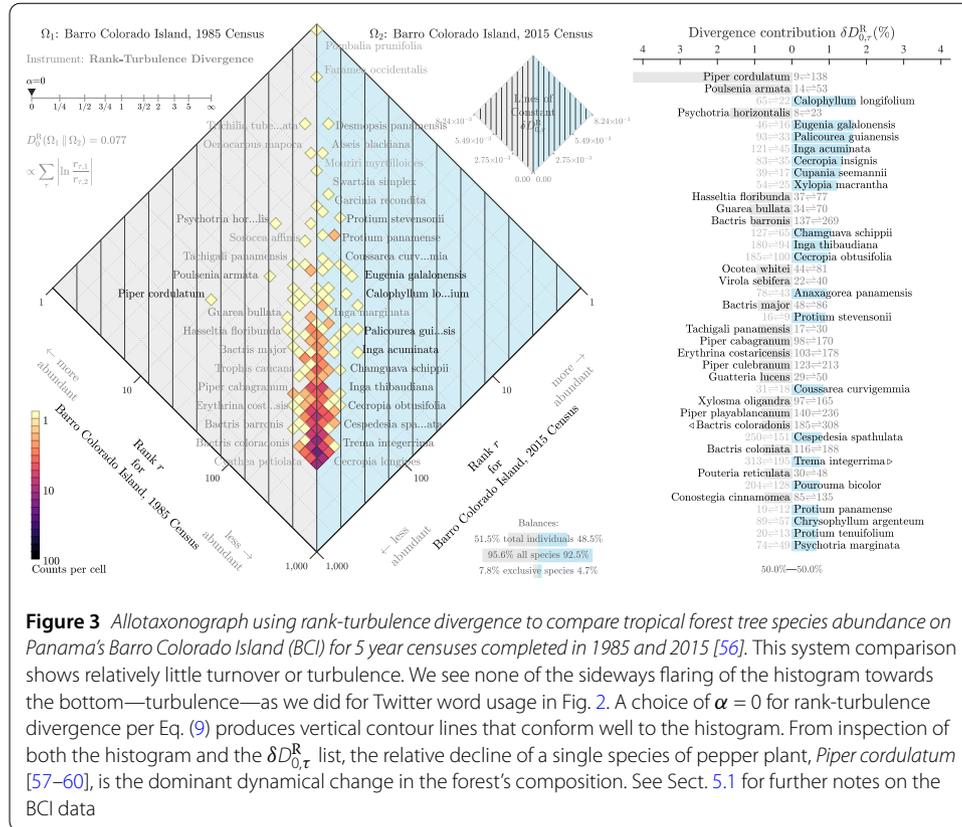


Figure 3 Allotaxonograph using rank-turbulence divergence to compare tropical forest tree species abundance on Panama's Barro Colorado Island (BCI) for 5 year censuses completed in 1985 and 2015 [56]. This system comparison shows relatively little turnover or turbulence. We see none of the sideways flaring of the histogram towards the bottom—turbulence—as we did for Twitter word usage in Fig. 2. A choice of $\alpha = 0$ for rank-turbulence divergence per Eq. (9) produces vertical contour lines that conform well to the histogram. From inspection of both the histogram and the $\delta D_{0,\tau}^R$ list, the relative decline of a single species of pepper plant, *Piper cordulatum* [57–60], is the dominant dynamical change in the forest's composition. See Sect. 5.1 for further notes on the BCI data

systems [62], the composition of an ecological system may change dramatically over a few decades. For this example from BCI, however, we see a system that is strongly durable in its component rankings.

We numerically compare the 1985 and 2015 distributions by applying rank-turbulence divergence with $\alpha = 0$, finding $D_0^R(R_1 \parallel R_2) = 0.077$. By inspection, we choose $\alpha = 0$ here because of the match of the histogram with the verticality of the contour lines (we address optimal selection of α in our concluding remarks). The nature of the BCI example affords us an opportunity to demonstrate the limit of $\alpha = 0$ for allotaxometry, and is a secondary reason for including an example from ecology.

In Flipbook S4, we show how the allotaxonomer performs with α varying away from 0 to ∞ . The visual match on the contour lines continuously degrades.

The BCI example's histogram is far from what we would expect of randomized systems (Fig. 11). To see how RTD quantifies the difference between two observed systems and then between randomized versions of these systems, we construct a set of pairs of randomized systems, measuring rank-turbulence divergence for each. We do this by randomly permuting the species names within each system while leaving species counts fixed, thereby keeping the size-rank distributions the same. We can perform such a randomization for any system-system comparison (and we do so below again for baby names).

We denote the average randomized divergence for two rankings R_1 and R_2 as:

$$D_{\alpha;\text{rand}}^R(R_1 \parallel R_2). \tag{13}$$

For the BCI example, we find that the score $D_0^R(R_1 \| R_2) = 0.077$ is well short of the randomized equivalent of $D_{0,\text{rand}}^R(R_1 \| R_2) = 0.376$ (average of 100 randomizations; standard deviation $\sigma = 0.012$).

Per Eq. (9), the contribution to overall divergence by changes in species abundance follows a log-ratio of ranks: $|\ln r_{\tau,1}/r_{\tau,2}|$. The contour lines for constant $\delta D_{0,\tau}^R$ accord with the histogram's form. From the histogram and $\delta D_{0,\tau}^R$ list, we see one species of pepper plant—*Piper cordulatum* [57–60]—stands out, having diminished markedly in relative abundance, dropping from $r_1 = 9$ to $r_2 = 138$. Two other species that have dropped in relative abundance feature in the top 4 of the $\delta D_{0,\tau}^R$ list: *Polsenia armata* ($r_1 = 14$ to $r_2 = 53$) and *Psychotria horizontalis* ($r_1 = 8$ to $r_2 = 23$).

Per the balance indicators, we see that the total number of individuals in each year's census is roughly the same (51.5% and 48.5%), that most types for both years appear in each system (95.6% and 92.5%), and that relatively few types are exclusive to each year (7.8% and 4.7%). Only two year-exclusive species make the top 40 for $\delta D_{0,\tau}^R$ contributions: *Bactris coloradonis* (1985 only) and *Trema integerrima* (2015 only). Regarding changes in overall diversity, we see that the loss of *Piper cordulatum* has not been to the gain of a single species—there is no one species on the right of the histogram with a distinctly high $\delta D_{0,\tau}^R$. Of the top 10 species ranked by $\delta D_{0,\tau}^R$, 7 are species that have become relatively more abundant. For the top 40, the balance is 20 down and 20 up. Overall, our instrument's dashboard makes clear that there is a singular drop in *Piper cordulatum*'s ecological role amid incremental (and possibly also important) changes for other species, straightforwardly directing future research attention.

3.4 Baby names: example rank-turbulence divergence allotaxonograph for the limit of $\alpha = \infty$

For an example of where tuning rank-turbulence divergence's parameter α to the limit of ∞ is helpful, we explore the temporal evolution of US baby name popularity [63, 64]. Because of the richness of baby name trends, we will also show how the full range of α can be used to uncover cultural changes.

The dataset we use tabulates annual name frequencies running from 1880 to 2021. The dataset is derived from Social Security card applications which means it is (unsurprisingly) not an exact measurement of baby name frequencies, particularly for retroactive registrations for those born in the years before Social Security was enacted in 1935.

For privacy, there is a truncation instituted in the dataset, and only baby names for which there are 5 or more instances in a year are included. Our discussion and analysis below therefore carries the caveat that rare names are occluded from our view (for further details and limitations see Sect. 5.1).

Because we will favor brevity in our discussion, when we write, for example, “baby girl names for 1968”, we will mean “US baby girl names registered at least 5 times with Social Security in the year 1968.”

In Fig. 4, we use a rank-turbulence divergence graph with $\alpha = \infty$ to compare changes in baby name frequencies for girls born in the US in 1968 and girls born in the US in 2018, a 50 year gap. In Fig. 5, we present the corresponding allotaxonomic graph for boy names. In the Ancillary files, we provide Flipbooks with $\alpha = \infty$ showing half century changes for both girl and boy names starting in 1880 and moving forward in 5 year increments (Flipbooks S5 and S6), as well as Flipbooks for the same 1968–2018 comparison with α varying from 0 to ∞ (Flipbooks S7 and S8). For baby names, an interactive

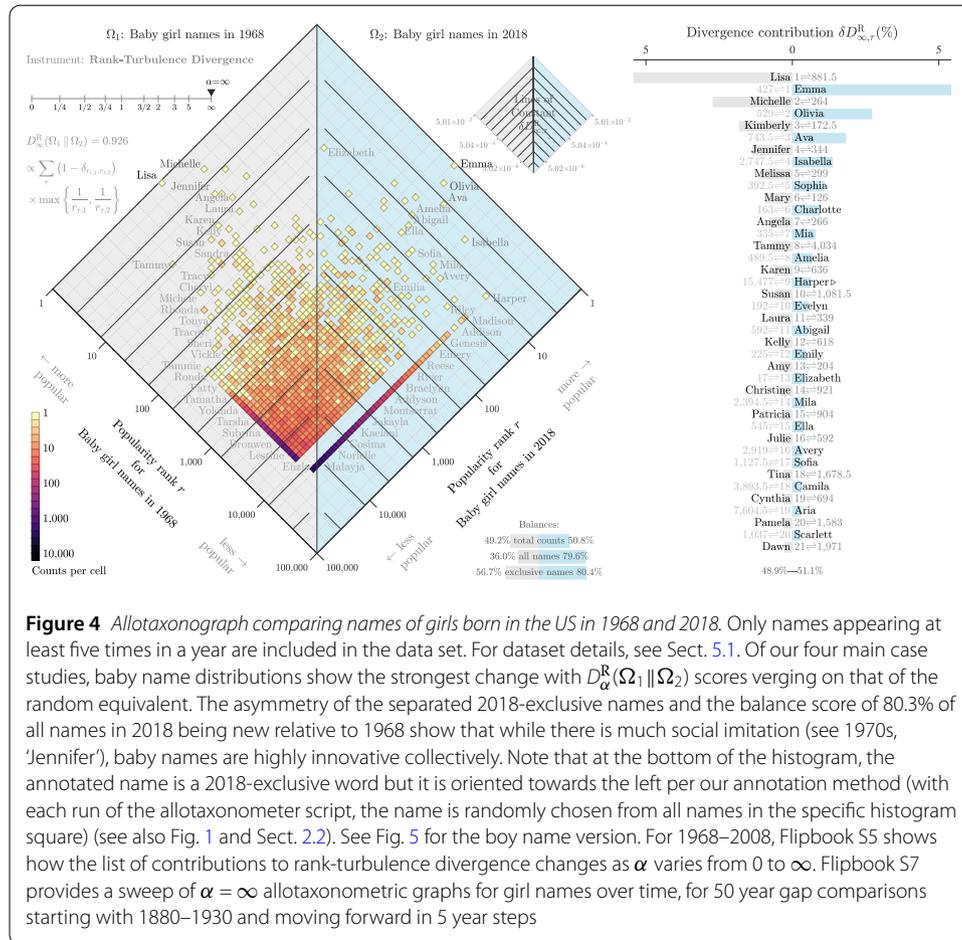


Figure 4 Allotaxonograph comparing names of girls born in the US in 1968 and 2018. Only names appearing at least five times in a year are included in the data set. For dataset details, see Sect. 5.1. Of our four main case studies, baby name distributions show the strongest change with $D_{\alpha}^R(\Omega_1 || \Omega_2)$ scores verging on that of the random equivalent. The asymmetry of the separated 2018-exclusive names and the balance score of 80.3% of all names in 2018 being new relative to 1968 show that while there is much social imitation (see 1970s, ‘Jennifer’), baby names are highly innovative collectively. Note that at the bottom of the histogram, the annotated name is a 2018-exclusive word but it is oriented towards the left per our annotation method (with each run of the allotaxonometer script, the name is randomly chosen from all names in the specific histogram square) (see also Fig. 1 and Sect. 2.2). See Fig. 5 for the boy name version. For 1968–2008, Flipbook S5 shows how the list of contributions to rank-turbulence divergence changes as α varies from 0 to ∞ . Flipbook S7 provides a sweep of $\alpha = \infty$ allotaxonometric graphs for girl names over time, for 50 year gap comparisons starting with 1880–1930 and moving forward in 5 year steps

version of the instrument would allow tunable α and the choice of years to be readily exploratory.

In contrast to the lexical turbulence of Twitter and the largely vertical form we saw for forest species counts, the histograms in Figs. 4 and 5 bear strong signatures of randomness and innovation.

First, as we saw in Fig. 1C, a random shuffling of ranked lists results in histograms predominantly weighted in the lower triangle of the plot. We see a strong imprint of this limiting case in Figs. 4 and 5, reflective of a great deal of cultural and societal change.

Second, we see dense exclusive-type lines at the base of both sides of the histograms in Figs. 4 and 5, the stamp of disjoint systems (Fig. 1D). The asymmetry of the histograms, with the separated exclusive-type line on the lower right, reflects the strong innovation of 2018 names relative to 1968. We note that the skew does not come from changes in system sizes as total numbers of births for the two years are comparable for girls and boys.

Overall, the turnover in baby names is stronger for girl names than boy names. We can gain a sense of this visually by observing that there is less flare to the left of the histogram for boy names relative to the histogram for girl names.

For girls, ranging from common 2018 names (‘Harper’, ‘Madison’, and ‘Addison’) down to rare names (e.g., ‘Kaisa’, ‘Akhar’, and ‘Hadly’), the 2018-exclusive names comprise 80.4%

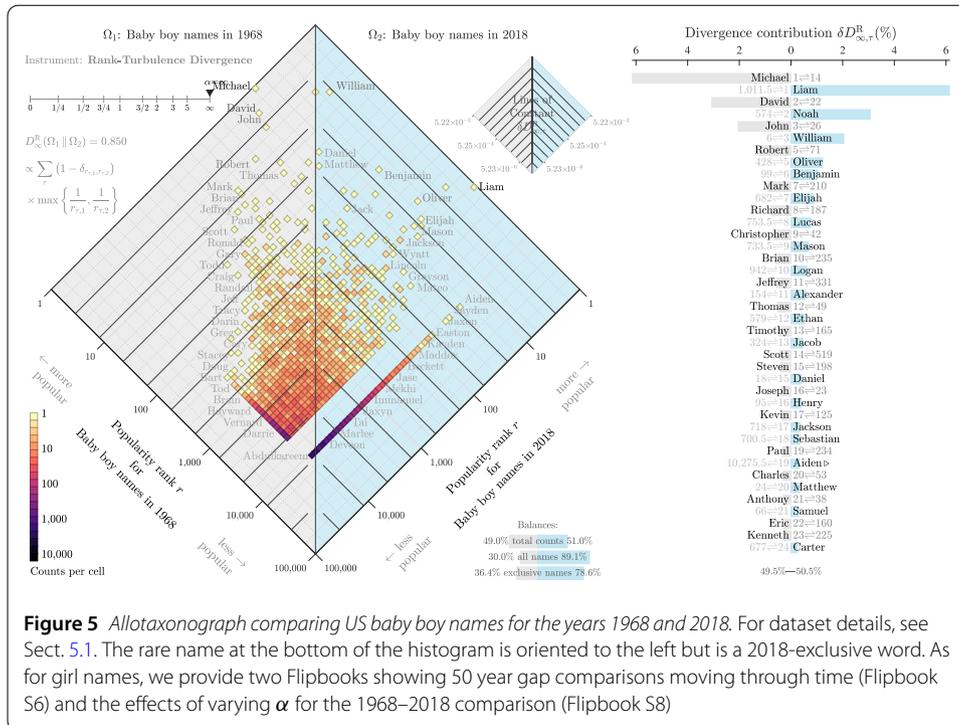


Figure 5 Allotaxonomograph comparing US baby boy names for the years 1968 and 2018. For dataset details, see Sect. 5.1. The rare name at the bottom of the histogram is oriented to the left but is a 2018-exclusive word. As for girl names, we provide two Flipbooks showing 50 year gap comparisons moving through time (Flipbook S6) and the effects of varying α for the 1968–2018 comparison (Flipbook S8)

of all names (14,563 of 18,115). For the smaller name base of boys, we see 78.6% (11,064 of 14,081) names are 2018-exclusive. Not registering above 5 counts in 1968 but widespread in 2018 are ‘Aiden’, ‘Jaxon’, and ‘Maddox’, and three 2018-exclusive but rare examples are ‘Kaston’, ‘Mak’, and ‘Cashis’.

While not separated because of the histogram’s cell sizes, the 1968-exclusive-type line is dense relative to the histogram body in both Figs. 4 and 5. We find 56.7% of all girl names (4,643 of 8,195) and 36.4% of all boy names (1,726 of 4,743) are 1968-exclusive names relative to 2018. A wide range of girl names that were popular in 1968 (‘Tammie’, ‘Ronda’, and ‘Patty’) as well as rare (‘Anmarie’ and ‘Adine’) have fallen out of favor by 2018. For boys, once-common ‘Bart’ and ‘Tod’ have dropped off the ledger. We also see apparent errors along the exclusive-type line for boy names in 1968 with ‘Gina’ (20 counts) and ‘Alicia’ (9 counts).

We emphasize that the balance indicators are for baby names appearing at least five times. For our present work, and in attempting to maintain uniformity across allotaxonomographs, we do not attempt to adjust for names appearing less than 5 times, though this would be possible for the topmost balance for total counts given we have that information separately. Clearly the balance values would shift if we had complete data sets for baby names; estimating errors for these estimates would be meaningful future work.

We note that the asymmetries of both histograms—their apparent right-side ‘heaviness’—are not due even in part to changes in overall numbers. Using total birth numbers (see Sect. 5.1), the total number of girl names recorded in 1968 and 2018 are comparable at 1,709,551 and 1,846,101 (7.99% increase); for boys, these numbers are 1,775,997 and 1,928,871 (8.61% increase). The number of year-exclusive names in the 1968 and 2018 are strikingly different however: 8,195 and 18,115 for girls (121% increase), and 4,743 and

14,081 for boys (197% increase). Two of the likely major factors which have lead to this explosion in name-space are immigration and a cultural shift towards parents creating novel names.

Using the overall birth numbers, we can also estimate the percentage of names absent from our dataset—those with less than 5 instances: 4.05% for 1968 and 8.08% for 2018 for girls, and 2.11% for 1968 and 6.07% for 2018 for boys. The 2018 size-rank distributions thus have heavier tails pointing once again to strong innovation.

The turnover in girl names results in a high rank-turbulence divergence value of $D_{\infty}^R(R_1 \| R_2) = 0.926$. For the same time frame comparison, boy names have a lesser but still high value of $D_{\infty}^R(R_1 \| R_2) = 0.850$. Both values are below but not far from the randomized equivalents with size-rank distributions held constant (as described in Sect. 3.3 for the BCI case): $D_{\infty, \text{rand}}^R(R_1 \| R_2) = 0.973$ and 0.966.

We turn to the overall orderings of $\delta D_{\infty, \tau}^R$ contributions for girls and boys, the ordered lists of Figs. 4 and 5.

In general, in the limit of $\alpha = \infty$, the contribution ordering will be an interleaving of types from both distributions. The ordering of types on each side of the list will match those of the separate size-rank distributions with the exception that all types that do not change rank will be absent. The interleaving is generally a simple back and forth sequence between the two systems but breaks whenever a rank is reached that is the lowest rank (largest value of r) for a specific type.

For girls in 1968 relative to 2018, we see the three medal places go to ‘Lisa’, ‘Michelle’, and ‘Kimberly’. In fourth, we have ‘Jennifer’, a name that would go on to be the most popular girl name in the US throughout the entire 1970s. In fifth is the once dominant ‘Mary’ which had held the number one position from 1880 almost entirely through to 1961 (‘Mary’ was second to ‘Linda’ for the years 1947–1952).

The dominance of the most popular girl name in 1968, ‘Lisa’, relative to 2018 is remarkable, carrying the top overall 1968 $\delta D_{\infty, \tau}^R$ contribution for all values of α . In Flipbook S7, we see that in dropping from $r = 1$ to $r = 888$, ‘Lisa’ is second in contribution for both 1968 and 2018 only for $\alpha = 0$ (first page) when we see ‘Harper’ take the top position. At this limit, order is by rank ratio and the above-the-rim elevation for ‘Harper’ from $r = 15,437$ to $r = 9$ is more than enough for the win.

On the other side, for 2018 relative to 1968, ‘Emma’ is the new ‘Lisa’, with ‘Olivia’ and ‘Ava’ in second and third for $\delta D_{\infty, \tau}^R$ contribution. In dialing α , Flipbook S7 shows that like ‘Lisa’, ‘Emma’ prevails above all other names except ‘Harper’ when $\alpha = 0$.

For boy names, the 1968 $\delta D_{\infty, \tau}^R$ side of the list is headed by ‘Michael’, ‘David’, ‘John’, and ‘Robert’ while for 2018, the top differential names are ‘Liam’, ‘Noah’, ‘William’, and ‘Oliver’. As we tune α down from ∞ to 0 (Flipbook S8), we see that ‘Liam’ has the top $\delta D_{\infty, \tau}^R$ contribution across all α , exceeding the ranges of ‘Lisa’ and ‘Emma’.

Of special note is the name ‘Elizabeth’ which stands out on the rank-rank histogram, well isolated in the upper triangle. We see that of all the top girl names in 1968, ‘Elizabeth’ alone has held its popularity. Flipbook S5, further shows that ‘Elizabeth’ maintains this isolated stability over decades. No standard divergence measure will highlight ‘Elizabeth’, inviting the development of a different class of measures that find anomalous rank-rank pairs.

While not to the degree of ‘Elizabeth’, there are two boy names that occupy a small hollowed-out region of rank-rank space in the histogram of Fig. 5: ‘James’ (steady at $r = 4$)

and ‘William’ (up from $r = 6$ to $r = 3$). As ‘Liam’ is an Irish variant on ‘William’, the latter effectively held the 1st and 3rd position in 2018.

For girl names compared with the α set to 0, the first page of Flipbook S5 shows that 1968 and 2018-exclusive names dominate the overall list. While ‘Lisa’ remains at the top, we then have ‘Tammy’, ‘Michele’, ‘Rhonda’, ‘Michelle’ and ‘Tammie’ as the 6 names from 1968 in the top 40 for $\delta D_{0,\tau}^R$ contributions. After ‘Harper’, the top 2018 names are ‘Madison’, ‘Isabella’, ‘Luna’, and ‘Layla’.

Using $\alpha = 0$ for boy names, we see in Flipbook S8 that only one name from 1968 makes the top 40 for $\delta D_{0,\tau}^R$ contributions: ‘Bart’. The top 40 list is otherwise all boy names from 2018, leading with ‘Liam’, ‘Aiden’, ‘Jayden’, ‘Noah’, and ‘Jaxon’.

Our allotaxonomic instrument also has the ability to uncover subsets of related types behaving in similar ways. For example, when tuning to $\alpha = 0$ (Flipbook S5), we see a raft of 2018-exclusive boy names ending in ‘-aden’, ‘-aiden’, and ‘-ayden’. Investigating further, we find 175 names appearing 5 or more times in 2018 that are exclusive to 2018 relative to 1968 and matching the regular expression `/[Aa][iy]*d+[aeoiuy]n+$/`. A selection of examples ranging from common to rare, highlighting variations on Brayden, are:

- ‘Aiden’ ($r = 19$)
- ‘Jayden’ (30),
- ‘Brayden’ (84),
- ‘Kayden’ (97),
- ‘Zayden’ (185.5),
- ‘Rayden’ (683),
- ‘Braydon’ (856),
- ‘Braiden’ (1239),
- ‘Bradyn’ (1936),
- ‘Grayden’ (1936),
- ‘Braydan’ (3534.5),
- ‘Braydin’ (3817.5),
- ‘Bladen’ (4974.5),
- ‘Blayden’ (5177),
- ‘Braidyn’ (5177),
- ‘Vayden’ (5870),
- ‘Braydyn’ (6873),
- ‘Wayden’ (7322),
- ‘Bradon’ (8434.5),
- ‘Slayden’ (8434.5),
- ‘Xzayden’ (10,155.5),
- ‘Blaiden’ (11,389.5),
- ‘Braydenn’ (13,042),
- and
- ‘Braidon’ (13,042).

For girl names, using a similar analysis for the ending -lyn, we find 535 names exclusive to 2018, the top four of which are:

- 'Adalynn' ($r = 108$),
- 'Adalyn' (144),
- 'Adelyn' (226),
- and
- 'Adelynn' (316).

There are 21 other names matching the pattern $/^A[aeiouy]*d+[aeiouy]l+[yi]+n+$/$.

There are 85 names exclusive to 1968 that are of the -lyn family led by

- 'Jerilyn' ($r = 1152.5$),
- 'Jacalyn' (1528.5),
- and
- 'Cherilyn' (1870.5),

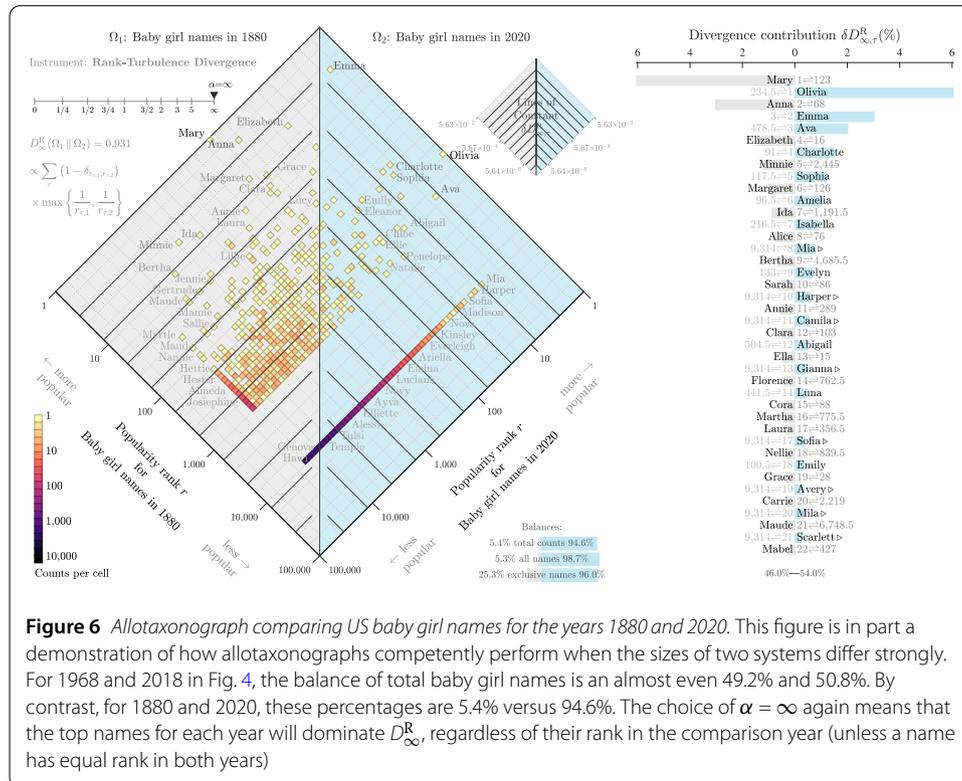
and 75 that appear in both 1968 and 2018 (e.g., 'Carolyn' and 'Evelyn').

These small interrogations of the data lead to larger questions which are beyond the scope of our work here. Are girl and boy names differently diverse? And how has the phonetic spread of names changed over time? A complete analysis could be performed by matching and grouping names based on spelling, syllables, and known variations.

To close out our study of baby names, we add two more allotaxonographs whose primary purpose is to show how our instrument performs when system sizes differ strongly. In Figs. 6 and 7, we compare US baby girl and boy names in 1880 and 2020, a 140 year gap.

We make some observations about balances, the rank-turbulence divergence scores, the rank-rank histograms, and the changes in naming from 1880 to 2020.

For the preceding allotaxonographs (Figs. 1–5), the largest difference for system sizes has been for Twitter in Fig. 1. The date 2016/11/09 carried 59.9% of all tweets from the



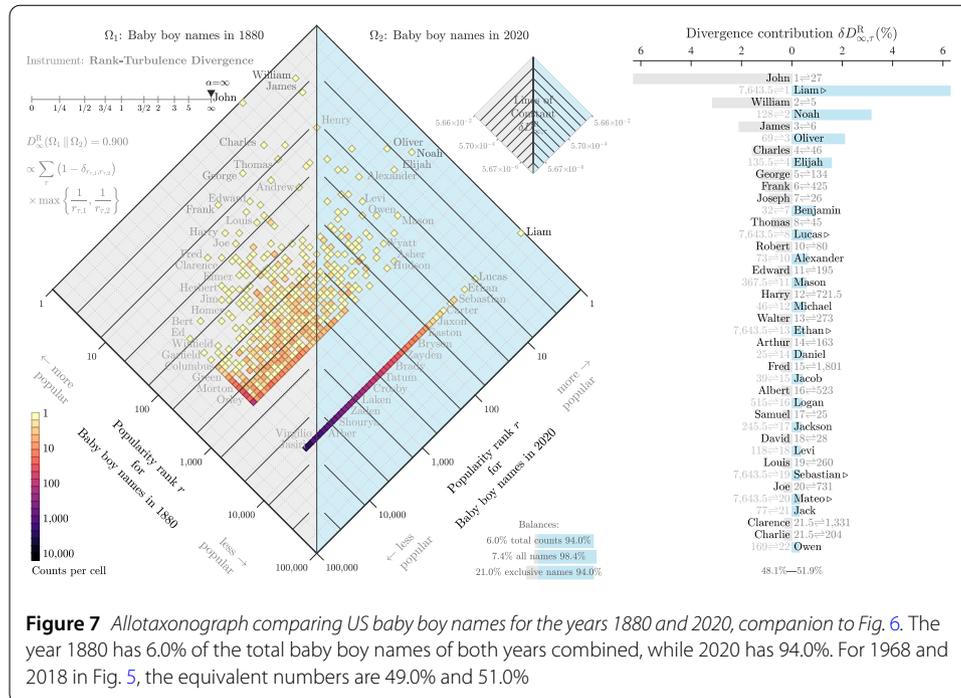


Figure 7 Allotaxonograph comparing US baby boy names for the years 1880 and 2020, companion to Fig. 6. The year 1880 has 6.0% of the total baby boy names of both years combined, while 2020 has 94.0%. For 1968 and 2018 in Fig. 5, the equivalent numbers are 49.0% and 51.0%

two dates combined, with the other 41.1% on 2017/08/18 (top balance bar, bottom right of the histogram).

By contrast, of the total number of baby girls born in 1880 and 2020, the years separately account for 5.4% and 94.6% respectively, a factor of roughly 17-fold (Fig. 6). For boys, these weights are similar at 6.0% and 94.0%, around 16-fold (Fig. 7).

Because of the large increase in registered babies being born, the two kinds of type balances are consequently more extreme. For example, of the combined types for the distinct baby girl names in 1880 and 2020, only 5.3% were used in 1880, while 98.7% were used in 2020. For exclusive types, 25.3% of 1880’s distinct baby girl names appeared only in 1880, while 96.0% of 2020’s were not used in 1880.

For baby girl names, the value of $D_\infty^R = 9.31$ for this 140 year comparison is slightly higher than that for the 50 year gap between 1968 and 2018, $D_\infty^R = 9.26$, while for boys the increase is from $D_\infty^R = 0.850$ to 0.900.

In general, the rank-rank histograms of these disparately sized systems will show a strongly separated, highly dense line corresponding to exclusive types on the side of the larger system. For both baby girl and boy names in 1880 and 2020, the separated line is around an order of magnitude from the main body of the histogram, and the component cells are high count ones. While this separation could occur for equal-sized systems if the type counts differ enough, the count density of the separated line will not be as strong. With familiarity, a glance at the balance bars will clarify these details.

Rank-turbulence divergence with $\alpha = \infty$ is a function only of the highest rank for each type (Eq. 11). As such, the main contributions for girls come from ‘Mary’ ($r = 1$ to 123) ‘Olivia’ ($r = 234.5$ to 1), while for boys the leaders are ‘John’ ($r = 1$ to 27) and ‘Liam’ ($r = 7643.5$ to 1, not used in 1880).

As we have reiterated, for evolving complex systems, allotaxonographs can help lead us to examine time series for individual types that occupy interesting locations in the rank-

rank histogram. For baby girl names, ‘Emma’ stands out as a name that was enormously popular in both 1880 ($r = 3$) and 2020 ($r = 2$). But the story for Emma proves to be akin to that of Vonnegut’s man-in-a-hole’s emotional arc [65–67]. Ranked third in 1880, ‘Emma’ dropped at a gradually increasing rate over the next 90 years to a stable set of low ranks in the 1970s—the decade of ‘Jennifer’—bottoming out at $r = 463$ in 1976. After first starting to revive in 1983, ‘Emma’ rapidly rose back to 4th in 2002 and stayed in the top 3 from 2003–2021, six times atop with $r = 1$.

3.5 Allotaxonomy of publicly traded US companies: stability, shocks, and errors

In Fig. 8, we show the rank-turbulence divergence graph comparing US company by market caps in the final quarter of 2007 with the final quarter of 2018 (for dataset description, see Sect. 5.1). The allotaxonograph is a blend of the two limiting cases of stability and change: The vertical line of matching systems and the ‘vee’ of disjoint systems (Figs. 1B and 1D). We choose $\alpha = 1/3$ for the rank-turbulence divergence instrument as the ordering of $\delta D_{1/3,\tau}^R$ values presents a mixture of high to low market cap (see below for more on this choice). In Flipbook S9, we show allotaxonographs for market cap comparisons for 6 year time gaps starting in 1995 and moving through to 2012.

Of the companies which both existed and reported market cap in both 2007 and 2018, we see a great deal of durability to their rankings. Somewhat more than what we see for species abundance numbers in Sect. 3.3, there are some notable movements in ranks. At the top of the rank-losing side of $\delta D_{1/3,\tau}^R$ list we see General Electric ($r = 2 \rightarrow 78$), Exxon

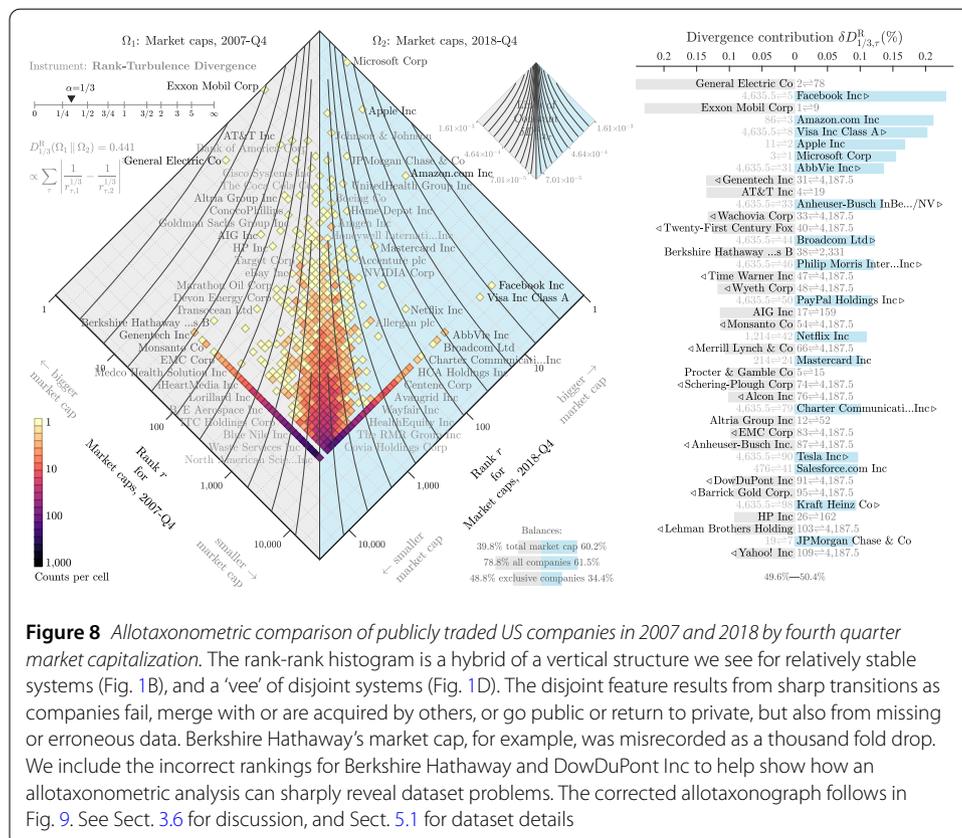


Figure 8 Allotaxonomic comparison of publicly traded US companies in 2007 and 2018 by fourth quarter market capitalization. The rank-rank histogram is a hybrid of a vertical structure we see for relatively stable systems (Fig. 1B), and a ‘vee’ of disjoint systems (Fig. 1D). The disjoint feature results from sharp transitions as companies fail, merge with or are acquired by others, or go public or return to private, but also from missing or erroneous data. Berkshire Hathaway’s market cap, for example, was misrecorded as a thousand fold drop. We include the incorrect rankings for Berkshire Hathaway and DowDuPont Inc to help show how an allotaxonomic analysis can sharply reveal dataset problems. The corrected allotaxonograph follows in Fig. 9. See Sect. 3.6 for discussion, and Sect. 5.1 for dataset details

Mobil (1 → 9), and AT&T (4 → 19). Berkshire Hathaway's apparent drop stems from a dataset error which we discuss below. On the right side for companies in existence in both 2007 and 2018, technology companies dominate: Amazon ($r = 86 \rightarrow 3$), Apple (11 → 2), Microsoft (3 → 1), and Netflix (1214 → 42).

Companies along the exclusive lines of the disjoint system 'vee' disappear and appear for a range of reasons. Mergers and acquisitions, companies being taken from public to private and vice versa, and outright failure all contribute to market cap comparisons having a disjoint aspect.

Looking through the 2007 exclusive companies on the histogram and the list (as indicated by the left triangle prefix), we see many companies that were acquired, with a few examples being Wachovia (bought by Wells Fargo in 2008), Genentech (bought by Roche in 2009), Time Warner (bought by Charter Communications in 2016), and Monsanto (bought by Bayer, 2018). We also find a few companies that failed with Lehman Brothers being a famous (or infamous) example from the 2007–2008 global financial crisis.

On the 2018 side, Visa and Facebook are the standout entrants. With respective initial public offerings (IPOs) in 2008 and 2012, we find they rank at $r = 5$ and 8 at the end of 2018. Visa's competitor Mastercard was already publicly traded in 2007, and ranks highly as well for $\alpha = 1/3$ ($r = 1214 \rightarrow 24$). AbbVie, Abbot Laboratories in 2013 ranks highest for pharmaceutical companies. The brewing company Anheuser-Busch InBev SA/NV formed in 2008 when Belgium's InBev purchased Anheuser-Busch.

The dataset for market caps does have some missing and erroneous data. DowDuPont's market cap for the last quarter of 2018 is absent and is consequently shown to have plummeted from a rank of $r = 91$ in 2007 to equal-to-last in 2018. Berkshire Hathaway's market cap is clearly misrecorded for the last three quarters of the dataset (apparently dropping from \$528.33B to \$0.34B at the end of 2018).

We take the opportunity to perform a small test of the sensitivity of rank-turbulence divergence by correcting the data for these two companies. For DowDuPont, with further sourcing, we find the year-end 2007 and 2018 market caps were reported as \$37.06B and \$121.34B, and for Berkshire Hathaway, \$149.56B and \$502.37B. Upon making these corrections, we first find again that $D_{1/3}^R = 0.411$, unchanged to three decimal places. In the corrected allotaxonograph (Fig. 9), Berkshire Hathaway's location shifts to the right side of the histogram ($r = 38 \rightarrow 5$) and is now listed as the 7th overall strongest contribution for $D_{1/3}^R$. DowDuPont no longer makes the top 40 of the list of contributions. While these two changes are dramatic, the remainder of the allotaxonograph remains essentially identical.

We have chosen to leave such errors in Fig. 8 to help (again) demonstrate the importance of using a rich, graphical allotaxonomic instrument. With a naive measurement of divergence, we would easily miss problematic data points. Evidently, and beyond our present paper's interests, for any further investigations, these two errors suggest that considerable effort should be made to further clean the market cap dataset

More generally, the specific form of the market cap histogram in Fig. 8 shows how we must take care when measuring divergences of any kind. The histogram's structure is not as simple as those for Twitter, species abundance, and baby names, and it would be problematic to allow for an unexamined, automated fitting of α for rank-turbulence divergence (or parameters of any other divergence).

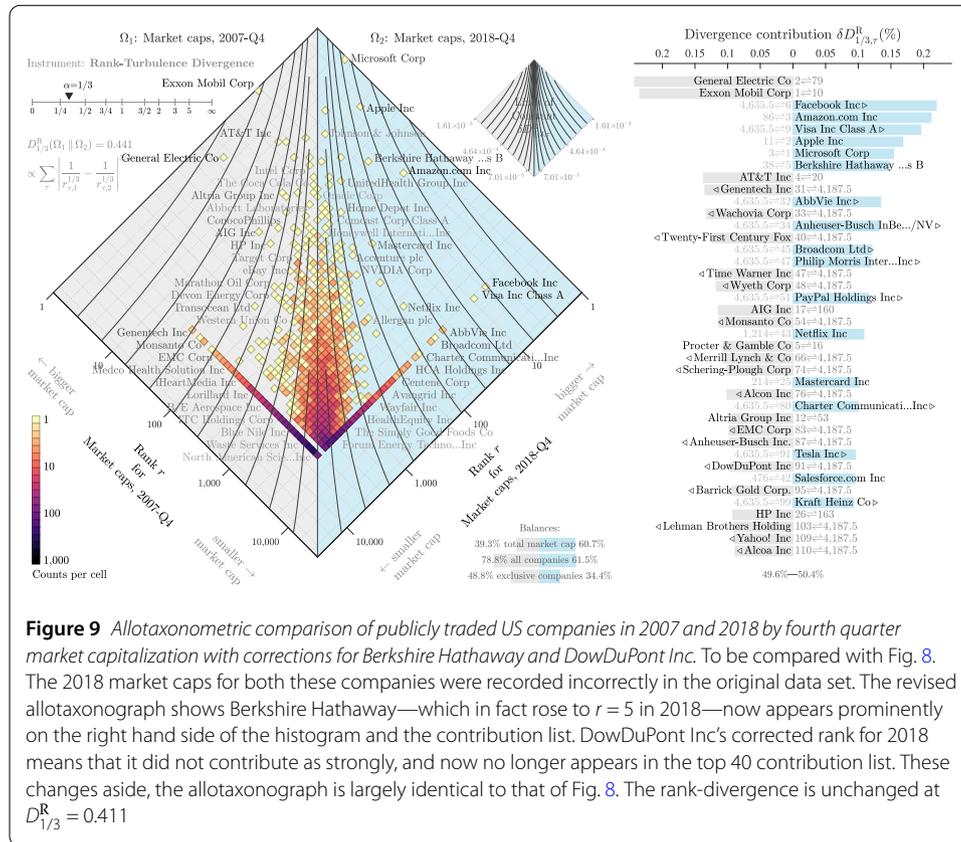


Figure 9 Allotaxonomic comparison of publicly traded US companies in 2007 and 2018 by fourth quarter market capitalization with corrections for Berkshire Hathaway and DowDuPont Inc. To be compared with Fig. 8. The 2018 market caps for both these companies were recorded incorrectly in the original data set. The revised allotaxonomogram shows Berkshire Hathaway—which in fact rose to $r = 5$ in 2018—now appears prominently on the right hand side of the histogram and the contribution list. DowDuPont Inc’s corrected rank for 2018 means that it did not contribute as strongly, and now no longer appears in the top 40 contribution list. These changes aside, the allotaxonomogram is largely identical to that of Fig. 8. The rank-divergence is unchanged at $D_{1/3}^R = 0.411$

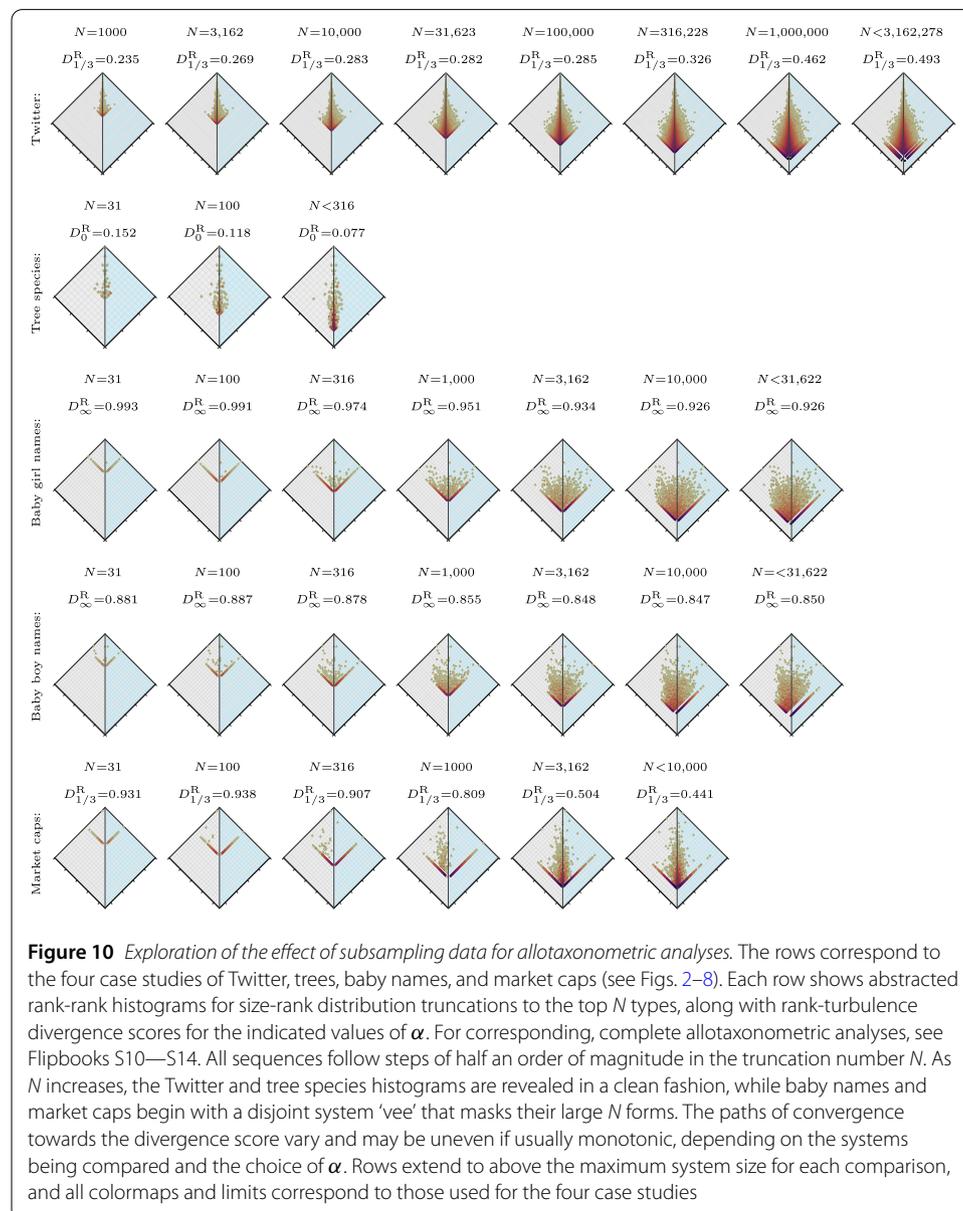
Given the composite form of the allotaxonomogram for market caps, an alternative treatment would be to separate out companies that appear in both systems from those companies that appear in only one year, the exclusive types. The enduring companies could be analyzed as a low-turbulent system on its own, and the companies exiting and entering as a disjoint system. A rank-based divergence instrument could be constructed that achieves this automatically, possibly returning a set of measurements that would capture that stable-shock balance we so clearly observe. Handling mergers, acquisitions, and partitionings of companies is also plausible and would require other kinds of elaboration of rank-turbulence divergence.

3.6 Truncation effects for rank-based allotaxonomograms

Truncation of a system’s size-rank distributions is a common if often overlooked problem [33, 68]. datasets may be curtailed for many reasons such as fundamental or cost-imposed measurement limits, data storage constraints, and privacy. Text corpora generate especially heavy-tailed distributions, with hapax legomena taking up roughly half of a text’s lexicon [14]. The Google Books n -gram corpus only includes n -grams which have appeared 40 or more times [69], excluding a vast number of rare n -grams. In our present work, we have already seen that for Twitter, our sample is approximately 10% of all tweets (with Twitter itself being a rather small subsample of all forms of human expression), and that baby names with counts of 4 or less are not made public for any censused population within the US. Limits to sampling in ecological systems can be severe—the Barro Colorado Island data is evidently not inclusive of all plant matter.

To investigate the problem of truncation, we explore our four case studies of Twitter, tree species, names, and companies by systematically limiting the observable components of each system. For each pair of systems, we take the top $N = 10^k$ ranked components where $k = 1.5, 2.0, 2.5, \dots$, stopping once we exceed the size of both systems. For each k , we generate the corresponding series of rank-turbulence divergence graphs, producing Flipbooks S10–S14. For a visual summary of these Flipbooks, we combine a subset of the (bare) rank-rank histograms to form Fig. 10.

The five rows of Fig. 10 correspond to our four case studies, with baby names contributing two rows. The first two examples of Twitter and tree species show a regular trend towards the full histogram. By contrast, baby names and market caps both appear to be disjoint when strong truncation is applied (small N). As N increases, the internal ran-



dom structure for baby names and the stable vertical structure for market caps start to be revealed by $N = 1000$.

For the computed values of rank-turbulence divergence, we use the same values of α as in our main studies: For Twitter, $\alpha = 1/3$, tree species, $\alpha = 0$, baby names, $\alpha = \infty$, and market caps $\alpha = 1/3$ (Figs. 2, 3, and 8).

In general, as N is increased, we see the main stories and patterns emerge. For Twitter, the election's imprint is clear for low N (Flipbook S10) with the texture of Charlottesville requiring more words to be included. The most dramatic changes in the lists of rank-turbulence divergence occur for baby names and market caps, as the system-exclusive types of these comparisons are masked for low N .

As a rough rule of thumb, the appearance of separated system-exclusive lines suggests that the underlying datasets are sufficiently rich enough to allow for a substantive allotaxonomic comparison. For the example of Twitter, and understanding that cell size matters, we see the separation occurs when N is moved from 100,000 to 1,000,000. We see no such separation for tree species however the vertical form representing stability unveils itself with increasing N in clear fashion.

We see that the values of D_α^R for the truncation sequences approach the 'true' value in largely monotonic, if different, ways. For the Twitter study, the value of $D_{1/3}^R$ is approached from below, deceptively exhibiting a flat section up to $N = 10^6$. The ecology example starts above and moves down towards the overall score of $D_0^R = 0.077$. Baby names and markets caps similarly both start above their respective overall scores for D_∞^R and $D_{1/3}^R$, and move downwards, though their scores for strong truncation are close to one as they appear to be disjoint systems. While the baby name scores drop slowly and not far (0.993 and 0.881 for 31 names for girls and boys down to 0.926 and 0.850 for all names), the market cap study only starts to gain more than the 'vee' shape when N is into the thousands. Because the market cap data comparison is a blend of large-scale turnovers around a relatively stable core, the drop is slow and then fast and further (0.931 for 31 corporations down to 0.441 for all corporations).

Our work aside, we expect any divergence measure will likely vary as orders of magnitude more data is included. And we add that in certain circumstances, choosing to truncate a data set may be a well justified treatment of data.

Finally, we note that while some form of truncation is a common measurement issue with real data for complex systems with many components, it is certainly not the only one. Exploring how other kinds of measurement errors affect rank-turbulence divergence would be a natural area of future work.

4 Guide to flipbooks

To help demonstrate rank-turbulence divergence as an allotaxonomic instrument, we have referenced a number of Flipbooks throughout the paper. We include these and other Flipbooks as supplementary information which can be found as part of our paper's online appendices at <http://compstorylab.org/allotaxonomy/flipbooks>.

Flipbooks are intended to be 'flipped through' back and forth using a PDF reader with the view set to 'single page' rather than continuous.

We list and briefly describe all Flipbooks here. Our flipbooks follow various formats which include: Comparisons of two systems with varying rank-turbulence divergence parameter α ; Comparisons of a series of system pairs, often through time; and Comparisons of systems with truncation applied (Sect. 3.6).

When α is varied the values are 0, $\frac{1}{12}$, $\frac{2}{12}$, $\frac{3}{12}$, $\frac{4}{12}$, $\frac{5}{12}$, $\frac{6}{12}$, $\frac{8}{12}$, 1, 2, 5, and ∞ .

Flipbook S1—Word use on Twitter: US Presidential Election (2016/11/09) versus the Charlottesville Unite the Right Rally (2017/08/13); Variation of α .

Flipbook S2—Word use on Twitter: US Presidential Election (2016/11/09) versus the Charlottesville Unite the Right Rally (2017/08/13); Variation of inclusion of retweets from 1% to 100%; $\alpha = 1/3$.

Flipbook S3—Word use on Twitter: Variation of time comparing 2019/01/04 going forward roughly logarithmically in number of days to a year ahead, 2020/01/03, the day of the assassination of Qasem Soleimani; $\alpha = 1/3$.

Flipbook S4—Tree species abundance on Barro Colorado Island: Fig. 3 with variation of α . The Flipbook shows how increasing α from 0 leads to an increasingly poor fit on the rank-rank histogram.

Flipbook S5—Baby girl names over time: Described in Sect. 3.4, comparisons of baby girl name distributions 50 years apart starting in 1880 and going forward in 5 year increments, with $\alpha = \infty$. Ends with Fig. 4.

Flipbook S6—Baby girl names, 1968–2018: Described in Sect. 3.4, shows effect of varying α , with Fig. 4 as the fifth page.

Flipbook S7—Baby boy names over time: Described in Sect. 3.4, comparisons of baby girl name distributions 50 years apart starting in 1880 and going forward in 5 year increments, with $\alpha = \infty$. Ends with Fig. 5.

Flipbook S8—Baby boy names, 1968–2018: Described in Sect. 3.4, shows effect of varying α , with Fig. 5 as the fifth page.

Flipbook S9—Market caps: Comparison of market caps for publicly traded companies in the fourth quarter six years apart, starting with 1995 versus 2001 and ending with 2012 versus 2018, and with α fixed at $1/3$.

Flipbook S10—Word use on Twitter, truncated: Full series of allotaxonographs corresponding to histograms of row 1 in Fig. 10 with $\alpha = 1/3$.

Flipbook S11—Tree species abundance, truncated: Full series of allotaxonographs corresponding to histograms of row 2 in Fig. 10 with $\alpha = 0$.

Flipbook S12—Baby girl names, truncated: Full series of allotaxonographs corresponding to histograms of row 3 in Fig. 10 with $\alpha = \infty$.

Flipbook S13—Baby boy names, truncated: Full series of allotaxonographs corresponding to histograms of row 4 in Fig. 10 with $\alpha = \infty$.

Flipbook S14—Market caps, truncated: Full series of allotaxonographs corresponding to histograms of row 5 in Fig. 10 with $\alpha = 1/3$.

Flipbook S15—Season total points scored by players in the National Basketball Association: Season to season comparison of total player points per season, $\alpha = 1/3$. The Flipbook starts with 1996–1997 versus 1997–1998 and ends in 2017–2018 versus 2018–2019. Rookies, retirements, injuries are all in evidence. For $\alpha = 1/3$, Carmelo Anthony in 2003–2004 has the strongest debut, just ahead of LeBron James in the same year. Overall, Dwyane Wade's 2008–2009 season produced the highest $\delta D_{1/3,\tau}^R$, moving from $r = 51$ to 1 over the previous year where he was limited in playing time with injuries. In 2008–2009, Wade's points per game of 30.2 would be the highest of his career but his team, the Miami Heat, would founder, achieving the worst record in the NBA.

Flipbook S16—Google Books, Fiction in 1948 versus 1987, 1-grams: The first of three Flipbooks exploring n -gram usage in books by varying α . We have elsewhere documented

the deeply problematic influence of scientific literature and individual books in Ref. [70], rendering the Google Books project unreliable, as is. Nevertheless, the Version 2 n -grams dataset for English fiction is worth exploring [27] with different instruments, and we are endeavoring separately to provide corrective measures. For 1948, we see characters and place names dominate, and these come from a few books (e.g., ‘Lanny Budd’, ‘Raintree County’). The 1987 side shows words that are not tied to specific books but rather cultural and temporal phenomena, as well as cruder language: ‘KGB’, ‘CIA’, ‘Vietnam’, ‘lesbian’, ‘television’, ‘computer’, and ‘fucking’. Tuning α towards ∞ , we can see pronouns changing slightly in rank with ‘her’ and ‘she’ elevating and ‘he’ and ‘his’ dropping.

Flipbook S17—Google Books, Fiction in 1948 versus 1987, 2-grams: For 2-grams, we again see character names dominate 1947 for low α (‘Sung Chiang’, ‘the Professor’), while ‘the CIA’ and ‘the KGB’ stand out for 1987. Increasing α brings in the same words as for 1-grams preceded by ‘the’ (‘the phone’, ‘the computer’). As $\alpha \rightarrow \infty$, bigrams with ‘not’ as part appear more strongly for 1987.

Flipbook S18—Google Books, Fiction in 1948 versus 1987, 3-grams: For 3-grams, while we still see characters and place names for 1947, we now have what we call ‘pathological hapax legomena’, words (or trigrams in this case) that occur once in many books. The 3-grams are all from standardized, legal-speak front matter coming from outside of the story: ‘change without notice’, ‘your local bookstore’, and ‘Cover art by’. A second kind of trigram that dominates appears to be one that appears as part of a book’s title printed on every page in the header or footer. As we increase α , we again see ‘not’ appearing in contributing 1987 trigrams. Because of the combinatorial explosion around words like ‘computer’ and ‘phone’, we no longer see them in the trigram lists. One upshot of this brief inspection of Google Books is to highlight the value of separately examining n -grams. We also note that the 3-gram example is our largest system-system comparison with system sizes on the order of 10^9 .

Flipbook S19—Harry Potter books, all 1-grams: Comparison of each Harry Potter book relative to all other books in the series combined, using $\alpha = 1/2$ (the single book is the right hand system, the merged set of 6 books the left system). Character names and major objects and places dominate, and the first book is most different from the others combined.

Flipbook S20—Harry Potter books, uncapitalized 1-grams: The same comparison as the previous Flipbook but now with all capitalized words excluded, as an example attempt to use a different lens on our allotaxonometer. Hagrid’s speech patterns in part separates Book 1 (‘yer’, ‘ter’), Book 3 has ‘rat’, ‘dementor’, and a relative abundance of em dashes (‘—’), Book 7 has ‘sword’, ‘wand’, and ‘goblin’. The dominant elements are things, places, and repeated actions (e.g., spells) and descriptors. To examine changes in functional word usage, which may reveal changes in Rowling’s writing, we would increase α as we did for Google Books. Again, we see the relative ease of taking subsets with ranks for allotaxometry.

Flipbook S21—Causes of Death in Hong Kong: Five year gap comparison of causes of death reported per year in Hong Kong, starting with 2001 versus 2006 and moving through to 2012 versus 2017. Overall, pneumonia is the leading cause of death. In the second half of the time frame, ‘kidney disease’ and ‘dementia’ stand out as becoming more prevalent. Deaths listed as due to heroin drop off markedly in 2012 and 2013 relative to 5 years before. We note that changes in diagnoses, practices, and categorization are all confounding issues.

Flipbook S22—Job titles: US job titles based on text analysis of online postings, 2007 compared with 2018; variation across three kinds of job categorization, from coarse- to fine-grained groupings, with suitable variation of α ($\alpha = 0$, $\alpha = 1/12$, and $\alpha = 1/3$).

5 Data and code

5.1 Datasets

Word usage on Twitter: Derived from an approximate 10% sample of Twitter collective by the Computational Story Lab from 2008 to 2020; English language detection performed per Ref. [45].

Species abundance on Barro Colorado Island: The dataset and its online repository for censuses taken over 35 years are described in Ref. [56].

Baby names: Data taken from Social Security Card applications as made public in 2022. (We caution that historical counts in this data set do change with each new release of baby name counts.) For each year from 1880–2021, the dataset includes all names which have 5 or more applications. Because Social Security Numbers were first issued at the end of 1936, there is a change in the dataset's nature as people moved from registering as adults to being solely registered at birth. While we use the dataset as is here, we note that there is a clear change in the male to female ratio with more boys being registered from 1940 onwards. Baby name dataset available here: <https://catalog.data.gov/dataset?tags=baby-names>. Separate dataset for total births available here: <https://www.ssa.gov/oact/babynames/numberUSbirths.html>.

Market cap data: The underlying dataset comprises 9322 US publicly traded companies that have been part of the S&P 500 at any point during the period of 1979–2018, or part of the Russell 3000 index from 1995 on. Data is available from Sibilis Research here: <http://sibilisresearch.com/data/us-equity-returns/>.

National Basketball Association: Dataset available here: <https://stats.nba.com/players/traditional/>.

Google Books n-grams: Version 2, English Fiction. We filtered the database to collect only *n*-grams containing simple latin characters. Dataset available here [69]: <https://books.google.com/ngrams>.

Causes of Death in Hong Kong: The dataset is described in Ref. [71–73] and has been well studied by others [74–79]. The dataset contains 892,055 death records between 1995 and 2017.

Job titles: Provided by Burning Glass, the dataset is derived from online postings (several million job openings per day, tens of thousands of sources). Raw listings are processed and categorized into two smaller taxonomies with natural-language algorithms.

5.2 Code

All scripts and documentation reside on Gitlab: <https://gitlab.com/compstorylab/allotaxonometer>.

For the present paper, we wrote the scripts to generate the allotaxonographs in MATLAB (Laboratory of the Matrix). We produced all figures and flipbooks using MATLAB Versions R2019b, R2020a, and R2021a. The core script is highly configurable and can be used to create a range of allotaxonographs as well as simple unlabeled rank-rank histograms. Instruments accommodated by the script include rank-turbulence divergence, probability-turbulence divergence [43], and generalized symmetric entropy divergence which includes Jensen-Shannon divergence as a special case.

Table 1 A selection of example system comparisons producing a range of $D_{\alpha}^R(\Omega_1 \parallel \Omega_2)$ values

Systems $\Omega^{(1)}$ and $\Omega^{(2)}$	Visualization/Section	α	$D_{\alpha}^R(\Omega_1 \parallel \Omega_2)$
Matching systems	Fig. 1B, Sect. 2.2	Any	0
Species, Barro Colorado Island	Fig. 3, Sect. 3.3	0	0.077
Causes of death, Hong Kong, 2012–2017	Flipbook S21, Sect. 4	1/3	0.213
Player season points, 2014–2015 vs 2015–2016	Flipbook S15, Sect. 4	1/3	0.279
Lowercase words, Harry Potter 7 vs 1–6	Flipbook S20, Sect. 4	1/2	0.308
Companies, 2007 vs 2018	Fig. 8, Sect. 3.5	1/3	0.441
Words on Twitter: 2016/11/09 vs 2017/08/13	Fig. 2, Sect. 3.1	1/3	0.493
Baby boy names, US, 1885 vs 1910	Sect. 3.4	∞	0.536
Baby girl names, US, 1900 vs 1925	Sect. 3.4	∞	0.631
Baby boy names, US, 1918 vs 1968	Flipbook S6, Sect. 3.4	∞	0.772
Baby boy names, US, 1968 vs 2018	Fig. 5, Sect. 3.4	∞	0.850
Baby girl names, US, 1918 vs 1968	Flipbook S5, Sect. 3.4	∞	0.887
Baby girl names, US, 1968 vs 2018	Fig. 4, Sect. 3.4	∞	0.926
Disjoint systems	Fig. 1D, Sect. 2.2	Any	1

6 Concluding remarks

6.1 Summary

Our goal has been to propose, advocate for, and contribute to a field of allotaxonomy: The measurement and visualization of detailed, type-level differences between complex systems. In the development of dynamic allotaxonomic dashboards, we have argued for a full embrace of complexity and stringent avoidance of falling into the trap of describing system differences solely by a single number.

In Sect. 1.3, we observed numerous benefits for using ranks: Widespread applicability beyond systems with type frequencies, probabilities, or rates; a natural handling of system-exclusive types by ranking them last; robustness of rank-based statistics; and the straightforward interpretability of ranked lists.

Focusing on systems with many components which can be ranked by some kind of well-defined size, we have created, tested, and explored rank-based allotaxographs built around our conception of a tunable rank-turbulence divergence. In Table 1, we collect a list of example system comparisons with $D_{\alpha}^R(\Omega_1 \parallel \Omega_2)$ ranging from 0 to 1.

At the core of rank-turbulence divergence in Eq. (7) is the interpretable difference of inverse powers of type ranks:

$$\left| \frac{1}{[r_{\tau,1}]^{\alpha}} - \frac{1}{[r_{\tau,2}]^{\alpha}} \right|. \tag{14}$$

As $\alpha \rightarrow 0$, the differences between ranks are contracted and low rank types become more salient. As $\alpha \rightarrow \infty$, rank discrepancies become more exacerbated, and the highest rank types dominate.

Narrowing our view to systems which afford frequencies of components, we find our directly tunable divergence appears to be far more general than many probability-based divergences, which are largely grouped around a few core structures. Per Ref. [31] and imposing the Zipf’s law ideal of $p = 1/r$, we see that $|r_{\tau,1}^{-1} - r_{\tau,2}^{-1}|$ is an abundant form. There are a few other variations including $\min(r_{\tau,1}, r_{\tau,2})$, and the Hellinger-like distance $|r_{\tau,1}^{-\frac{1}{2}} - r_{\tau,2}^{-\frac{1}{2}}|$. These three cases correspond to our rank-turbulence divergence with $\alpha = 1, \infty$, and $1/2$.

For the instrument’s integrity and power, we assert that the map and list should be bound together. While our allotaxonomic histograms give immediate stories from the automati-

cally labeled words along the fringes, the overall ordering of these words by some measure of importance is unclear. And in choosing to map a two-dimensional rank-rank histogram onto a single dimension—another ranked list—we remain mindful that we are discarding information. We suggest that, analogously, all cartograms would benefit from an associated ordered list and vice versa [10].

As we have stated, there is tendency across diverse fields towards creating single-number measurements of complex systems, and that this is especially problematic when heavy-tailed size-rank distributions are in evidence (e.g., the Gini coefficient). We have shown that even when single-number measures match for two systems, allotaxonographs using rank-turbulence divergence are able to reveal and make sense of the full variation between systems.

The four main case studies of Twitter, tree species, baby names, and companies have all provided rich and diverse examples of allotaxonomic comparisons. Our ability to readily analyze the effects of partially sampled data in Sect. 3.6 further showed the value of a rank-based approach. Drawing on our paper's preprint, we and others have also used allotaxonographs in a number of other papers [80–85].

With our supplementary Flipbooks, we have attempted to show the prospect for the building of online, interactive allotaxonographs. Being linear in nature, Flipbooks allow us to explore one dimension of variation at a time, and by design are built to be fixed rather than flexible. For baby names, for example, we would like to be able to interactively vary the years being compared as well as rank-turbulence divergence's α . For temporally evolving systems, an interactive allotaxonograph could be set to track a particular cohort of types or to automatically highlight those which make a dynamical transition of some prescribed kind.

There are many future research possibilities, both theoretical and applied, suggested or opened up by what we have developed here for rank-turbulence divergence and, more generally, for allotaxonomy.

6.2 Theoretical foundation and other allometric instruments

We have been pragmatic in our construction of rank-turbulence divergence, striving to build a functional tool first and foremost. A rigorous theoretical foundation might be possible for either our tool or an adjacent rank-based divergence. Staying on the functional side, variations on our divergence might be of use for some comparisons where no value of α makes for a good fit. As we noted for the case of market caps, a composite instrument that separates stable, enduring companies to those that exit or enter could be devised.

For systems with documented component probabilities or rates, we have also constructed a related probability-turbulence divergence. We explore the allotaxonomy of this divergence in [43], showing the instrument to be a generalization of a suite of well known probability-based divergences.

As we saw for the unusually durable popular name 'Elizabeth' in Fig. 4, there are components whose locations on allotaxonographs are not highlighted by standard conceptions of divergences, rank-based or otherwise. A completely distinct measure of importance could favor largely isolated rank-rank pairings on the rank-rank histogram. Given that the measure would have to be sufficiently sophisticated to accommodate the possibility that a small cluster of related types might be near each other (e.g., 'Lady' and 'Gaga'), yet otherwise be distinct, the application of some basic kind of cluster analysis would offer a starting point.

6.3 Determination of α

In our initial work, we have made the choice of the tuning parameter for rank-turbulence divergence, α , a visually guided one. The user gains much from inspecting the rank-rank histogram alone, and, in our experience, is then readily able to choose an α for which the allotaxonomic contour lines best match the form of the histogram. A visually guided choice will be sufficient in cases of comparing two or a small number of systems.

When rank turbulence presents as a scaling law—which is regularly the case for text corpora (e.g., Twitter, books)—we would want to be able to determine an optimal α . While for generalized entropy approaches for single systems, the limit of linear scaling and Shannon's entropy demarcate the boundary between accentuating the common or the rare [8, 34, 54, 55], we have found that for system comparisons, the optimal value of α , if it exists, is dependent on the pair of systems being compared—there is no universal value.

We have left open the possibility of an analytic connection between the rank-turbulence scaling described at the end of Sect. 1.2, and, to the extent that well-defined scaling is present, with an optimal α for rank-turbulence divergence.

Even with an optimization method for determining α , we urge readers to always look at the visuals provided by our allotaxonomographs—the maps—for confirmation of fit.

6.4 Rank energy

For another direction, we venture that a kind of 'rank energy' interpretation might be possible. Working from the idealized Zipf's law relationship of $p \sim r^{-1}$, we would have

$$p^\alpha \sim 1/r^\alpha = \exp\{-\alpha E/T\} = \exp\{-E/T'\}, \quad (15)$$

where $E = T \ln r$ is an energy associated with rank r and temperature T , and T' an effective temperature. When $T' \rightarrow 0$, high ranked types prevail, while when $T' \rightarrow \infty$, all types move towards being weighted equally, independent of rank.

6.5 Type calculus

Identifying and quantifying change is fundamental to any form of scientific analysis (and life itself). Allotaxonomy may be viewed as part of a larger analytic framework of 'lexical calculus' and, more generally, 'type calculus.' By lexical calculus, we mean the measurement of changes in properties of large-scale texts, and the demonstration of how individual word usage contributes to such changes through word shift graphs [86–90]. Expanding to complex systems comprising many types (which we would likely still denote by words), we would have a corresponding type calculus (e.g., baby names, companies, species). Simply measuring overall numeric changes in, say, entropy between two complex systems is grossly insufficient for understanding how systems may be differentially configured. We must always look at the words (or types).

6.6 Final remarks

We close with the observation that in terms of applications, any comparison of complex systems entailing a broad array of components would be fair game. A few examples would be sales of anything (e.g., Amazon's sales from week to week), crime rates, country exports, sites visited or searched for online, medical condition prevalences, rankings in sports, music popularity, and markets of all kinds. And while our focus has been on comparing sys-

tems at the level of components, changes in system structure, e.g., complex networks, could also be readily explored with the same rank-turbulence divergence instrument.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1140/epjds/s13688-023-00400-x>.

Additional file 1. (ZIP 238.7 MB)

Acknowledgements

The authors are grateful for support furnished by MassMutual and Google, and the computational facilities provided by the Vermont Advanced Computing Center. The authors are grateful for conversations with R. Gallagher, L. Mitchell, James O'Dwyer, and J. Weitz.

Funding

The authors acknowledge support from MassMutual, Google, and National Science Foundation grant #2242829.

Availability of data and materials

Code to produce the graphs is available in the paper's Gitlab repository.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author contributions

PSD conceived the idea; PSD, JRM, MVR, TA, JLA, DRD, AJR, and CMD collected and analyzed data; PSD wrote the figure code and the manuscript; TA, TJG, JRM, MVA, JLA, MRF, AJR, and CMD edited the manuscript. All authors read and approved the final manuscript.

Author details

¹Computational Story Lab, Vermont Complex Systems Center, MassMutual Center of Excellence for Complex Systems and Data Science, Vermont Advanced Computing Center, University of Vermont, Burlington, VT 05405, USA. ²Department of Computer Science, University of Vermont, Burlington, VT 05405, USA. ³Santa Fe Institute, 1399 Hyde Park Rd, Santa Fe, NM 87501, USA. ⁴Department of Mathematics & Statistics, University of Vermont, Burlington, VT 05405, USA. ⁵Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. ⁶MassMutual Data Science, Amherst, MA 01002, USA.

Received: 29 November 2020 Accepted: 16 June 2023 Published online: 19 September 2023

References

1. Dyson F (1993) George Green and physics. *Phys World* 6(8):33
2. Borland D, Wang W, Wang J, Shrestha J, Gotz D (2019) Selection bias tracking and detailed subset comparison for high-dimensional data. Available online at <https://arxiv.org/abs/1906.07625>
3. Diamond JM (1997) *Guns, germs, and steel*. Norton, New York
4. Turchin P, Currie TE, Whitehouse H, François P, Feeney K, Mullins D, Hoyer D, Collins C, Grohmann S, Savage P et al (2018) Quantitative historical analysis uncovers a single dimension of complexity that structures global variation in human social organization. *Proc Natl Acad Sci* 115:E144–E151
5. Strang G (2009) *Introduction to linear algebra*, 4th edn. Cambridge Wellesley Press, Wellesley
6. Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27(379–423):623–656
7. Shannon CE (1956) The bandwagon. *IRE Trans Inf Theory* 2(1):3
8. Jost L (2006) Entropy and diversity. *Oikos* 113:363–375
9. Dodds PS, Alshaabi T, Fudolig MI, Zimmerman JW, Lovato J, Beaulieu S, Minot JR, Arnold MV, Reagan AJ, Danforth CM (2021) Ousiometrics and telegnomics: the essence of meaning conforms to a two-dimensional powerful-weak and dangerous-safe framework with diverse corpora presenting a safety bias. Available online at <https://arxiv.org/abs/2110.06847>
10. Alajajian SE, Williams JR, Reagan AJ, Alajajian SC, Frank MR, Mitchell L, Lahne J, Danforth CM, Dodds PS (2017) The lexicocalorimeter: gauging public health through caloric input and output on social media. *PLoS ONE* 12:e0168893. arXiv version available at <http://arxiv.org/abs/1507.05098>
11. Peirce CSS (1906) Prolegomena to an apology for pragmatism. *Monist* 16(4):492–546
12. Zipf GK (1949) *Human behaviour and the principle of least-effort*. Addison-Wesley, Cambridge
13. Stigler SM (1980) Stigler's law of eponymy. *Trans N Y Acad Sci* 39:147–157
14. Simon HA (1955) On a class of skew distribution functions. *Biometrika* 42:425–440
15. Newman MEJ (2005) Power laws, Pareto distributions and Zipf's law. *Contemp Phys* 46:323–351
16. Coromina-Murtra B, Solé R (2010) Universality of Zipf's law. *Phys Rev E* 82:011102
17. Gerlach M, Font-Clos F, Altmann EG (2016) Similarity of symbol frequency distributions with heavy tails. *Phys Rev X* 6:021009

18. Williams JR, Lessard PR, Desu S, Clark EM, Bagrow JP, Danforth CM, Dodds PS (2015) Zipf's law holds for phrases, not words. *Nat Sci Rep* 5:12209
19. Barabási A-L, Albert R (1999) Emergence of scaling in random networks. *Science* 286:509–511
20. Axtell R (2001) Zipf distribution of U.S. firm sizes. *Science* 293(5536):1818–1820
21. Maillart T, Sornette D, Spaeth S, von Krogh G (2008) Empirical tests of Zipf's law mechanism in open source Linux distribution. *Phys Rev Lett* 101(21):218701
22. Miller GA (1957) Some effects of intermittent silence. *Am J Psychol* 70:311–314
23. Miller GA (1965) Introduction to reprint of G. K. Zipf's "The psycho-biology of language". MIT Press, Cambridge
24. Ferrer-i-Cancho R, Elvevåg B (2010) Random texts do not exhibit the real Zipf's law-like rank distribution. *PLoS ONE* 5:e9411
25. Mandelbrot BB (1953) An informational theory of the statistical structure of languages. In: Jackson W (ed) *Communication theory*. Butterworth, Woburn, pp 486–502
26. Dodds PS, Dewhurst DR, Hazlehurst FF, Van Oort CM, Mitchell L, Reagan AJ, Williams JR, Danforth CM (2017) Simon's fundamental rich-get-richer model entails a dominant first-mover advantage. *Phys Rev E* 95:052301
27. Pechenick EA, Danforth CM, Dodds PS (2017) Is language evolution grinding to a halt? The scaling of lexical turbulence in English fiction suggests it is not. *J Comput Sci* 21:24–37
28. Ferrer-i-Cancho R, Solé RV (2001) Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *J Quant Linguist* 8(3):165–173
29. Williams JR, Bagrow JP, Danforth CM, Dodds PS (2015) Text mixing shapes the anatomy of rank-frequency distributions. *Phys Rev E* 91:052811
30. Deza M-M, Deza E (2006) *Dictionary of distances*. Elsevier, Amsterdam
31. Cha S-H (2007) Comprehensive survey on distance/similarity measures between probability density functions. *Int J Math Models Methods Appl Sci* 1:300–307
32. Cichocki A, Amari S-I (2010) Families of Alpha- Beta- and Gamma- divergences: flexible and robust measures of similarities. *Entropy* 12:1532–1568
33. Haegeman B, Hamelin J, Moriarty J, Neal P, Dushoff J, Weitz JS (2013) Robust estimation of microbial diversity in theory and in practice. *ISME J* 7:1092
34. Hill MO (1973) Diversity and evenness: a unifying notation and its consequences. *Ecology* 54(2):427–432
35. Gotelli NJ, Colwell RK (2011) Estimating species richness. *Biol Divers Front Meas Assess* 12:39–54
36. Chao A, Gotelli NJ, Hsieh T, Sander EL, Ma K, Colwell RK, Ellison AM (2014) Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecol Monogr* 84:45–67
37. Merritt S, Clauset A (2014) Scoring dynamics across professional team sports: tempo, balance and predictability. *EPJ Data Sci* 3:4
38. Clauset A, Kogan M, Redner S (2015) Safe leads and lead changes in competitive team sports. *Phys Rev E* 91:062815
39. Kiley DP, Reagan AJ, Mitchell L, Danforth CM, Dodds PS (2016) The game story space of professional sports: Australian rules football. *Phys Rev E* 93:052314
40. Fagin R, Kumar R, Sivakumar D (2003) Comparing top *k* lists. *SIAM J Discrete Math* 17:134–160
41. Bar-Ilan J, Mat-Hassan M, Levene M (2006) Methods for comparing rankings of search engine results. *Comput Netw* 50(10):1448–1463
42. Webber W, Moffat A, Zobel J (2010) A similarity measure for indefinite rankings. *ACM Trans Inf Syst* 28:1–38
43. Dodds PS, Minot JR, Arnold MV, Alshaabi T, Adams JL, Dewhurst DR, Reagan AJ, Danforth CM (2020) Probability-turbulence divergence: a tunable allotaxonomic instrument for comparing heavy-tailed categorical distributions. Available online at <http://arxiv.org/abs/2008.13078>
44. Gray TJ, Danforth CM, Dodds PS (2020) Hahahahaha, duuuuuude, yeeesssss!: a two-parameter characterization of stretchable words and the dynamics of mistypings and misspellings. *PLoS ONE* 15:e0232938. Available online at <https://arxiv.org/abs/1907.03920>
45. Alshaabi T, Dewhurst DR, Minot JR, Arnold MV, Adams JL, Danforth CM, Dodds PS (2020) The growing amplification of social media: measuring temporal and social contagion dynamics for over 150 languages on Twitter for 2009–2020. *EPJ Data Sci* 10:15
46. Liu Y, Heer J (2018) Somewhere over the rainbow: an empirical assessment of quantitative colormaps. In: *Proceedings of the 2018 CHI conference on human factors in computing systems*. ACM, New York, p 598
47. Monroe BL, Colaresi MP, Quinn KM (2008) Fightin' words: lexical feature selection and evaluation for identifying the content of political conflict. *Polit Anal* 16:372–403
48. Kessler JS (2017) Scattertext: a browser-based tool for visualizing how corpora differ. *arXiv preprint*. [arXiv:1703.00565](https://arxiv.org/abs/1703.00565)
49. Bergstrom CT, West JD (2018) Why scatter plots suggest causality, and what we can do about it. Available online at <https://arxiv.org/abs/1809.09328>
50. Sonyeondan B (Korean: 방탄소년단; Hanja: 防彈少年團) meaning "Bulletproof Boy Scouts." In 2017, the band formally acknowledged the backronym "Beyond the Scene" as a secondary official name
51. Dodds PS, Minot JR, Arnold MV, Alshaabi T, Adams JL, Dewhurst DR, Reagan AJ, Danforth CM (2022) Fame and ultrafame: measuring and comparing daily levels of 'being talked about' for United States' presidents, their rivals, God, countries, and K-pop. *Journal of Quantitative Description: Digital Media* 2. Available online at <https://arxiv.org/abs/1910.00149>
52. Identity Evropa. Wikipedia (2019). https://en.wikipedia.org/w/index.php?title=Identity_Evropa&oldid=934670726. Accessed on 2020/01/28
53. Rényi A (1961) On measures of entropy and information. In: *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1: contributions to the theory of statistics
54. Tsallis C (2001) I. Nonextensive statistical mechanics and thermodynamics: historical background and present status. In: *Nonextensive statistical mechanics and its applications*. Springer, Berlin, pp 3–98
55. Keylock CJ (2005) Simpson diversity and the Shannon–Wiener index as special cases of a generalized entropy. *Oikos* 109:203–207
56. Condit R et al (2019) Complete data from the Barro Colorado 50-ha plot: 423617 trees, 35 years, v3, DataONE Dash, Dataset

57. Trelease W (1927) The Piperaceae of Panama. Systematic plant studies: mainly tropical America
58. Standley PC (1927) The flora of Barro Colorado Island, Panama. Smithsonian miscellaneous collections
59. Thies W, Kalko EKV (2004) Phenology of Neotropical pepper plants (Piperaceae) and their association with their main dispersers, two short-tailed fruit bats, *Carollia perspicillata* and *C. castanea* (Phyllostomidae). *Oikos* 104(2):362–376
60. Andrade TY, Thies W, Rogeri PK, Kalko EKV, Mello MAR (2013) Hierarchical fruit selection by Neotropical leaf-nosed bats (Chiroptera: Phyllostomidae). *J Mammal* 94(5):1094–1101
61. Condit R, Ashton P, Baker P, Sarayudh B, Gunatilleke S, Gunatilleke N, Hubbell S, Foster R, Itoh A, LaFrankie J, Lee H, Losos E, Manokaran N, Sukumar R, Yamakura T (2000) Spatial patterns in the distribution of tropical tree species. *Science* 288:1414–1418
62. Strogatz SH (1994) Nonlinear dynamics and chaos. Addison-Wesley, Reading
63. Hahn MW, Bentley RA (2003) Drift as a mechanism for cultural change: an example from baby names. *Proc R Soc Lond B, Biol Sci* 270:S120–S123
64. Wattenberg M (2005) Baby names, visualization, and social data analysis. In: IEEE symposium on information visualization, 2005. INFOVIS 2005. IEEE, Los Alamitos, pp 1–7
65. Kurt Vonnegut on the shapes of stories (2010) <https://www.youtube.com/watch?v=oP3c1h8v2ZQ>, accessed May 15, 2014
66. Vonnegut K Jr (2005) A man without a country. Seven Stories Press, New York
67. Reagan AJ, Mitchell L, Danforth CM, Dodds PS (2016) The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Sci* 5:31. Available at <http://arxiv.org/abs/1606.06820>
68. Koplenig A, Wolfer S, Müller-Spitzer C (2019) Studying lexical dynamics and language change via generalized entropies: the problem of sample size. *Entropy* 21:464
69. Michel J-B, Shen YK, Aiden AP, Veres A, Gray MK, The Google Books Team, Pickett JP, Hoiberg D, Clancy D, Norvig P, Orwant J, Pinker S, Nowak MA, Lieberman EA (2011) Quantitative analysis of culture using millions of digitized books. *Sci Mag* 331:176–182
70. Pechenick EA, Danforth CM, Dodds PS (2015) Characterizing the Google books corpus: strong limits to inferences of socio-cultural and linguistic evolution. *PLoS ONE* 10:e0137041
71. Micro-data set of known and registered deaths (2018) https://www.censtatd.gov.hk/service_desk/list/microdata/index.jsp Data retrieved from Census and Statistics Department of Hong Kong
72. District and constituency area (2016) <https://www.bycensus2016.gov.hk/en/bc-dp.html>. Data retrieved from Census and Statistics Department of Hong Kong
73. Department of Hong Tertiary planning units. <https://www.bycensus2016.gov.hk/en/bc-dp-tpu.html>, 2016. Data retrieved from Census and Statistics Kong
74. Wong C-M, Ma S, Hedley AJ, Lam T-H (2001) Effect of air pollution on daily mortality in Hong Kong. *Environ Health Perspect* 109(4):335–340
75. Lam T-H, Ho S-Y, Hedley AJ, Mak K-H, Leung GM (2004) Leisure time physical activity and mortality in Hong Kong: case-control study of all adult deaths in 1998. *Ann Epidemiol* 14(6):391–398
76. Ou C-Q, Hedley AJ, Chung RY, Thach T-Q, Chau Y-K, Chan K-P, Yang L, Ho S-Y, Wong C-M, Lam T-H (2008) Socioeconomic disparities in air pollution-associated mortality. *Environ Res* 107:237–244
77. Qiu H, Tian L, Ho K-F, Pun VC, Wang X, Ignatius T (2015) Air pollution and mortality: effect modification by personal characteristics and specific cause of death in a case-only study. *Environ Pollut* 199:192–197
78. Wong IO, Schooling C, Cowling BJ, Leung GM (2015) Breast cancer incidence and mortality in a transitioning Chinese population: current and future trends. *Br J Cancer* 112(1):167–170
79. Wu P, Presanis AM, Bond HS, Lau EH, Fang VJ, Cowling BJ (2017) A joint analysis of influenza-associated hospitalizations and mortality in Hong Kong, 1998–2013. *Sci Rep* 7(1):929
80. Gothard K, Dewhurst DR, Minot JA, Adams JL, 5-Danforth CM, Dodds PS (2021) The incel lexicon: deciphering the emergent cryptolect of a global misogynistic community. Available online at <https://arxiv.org/abs/2105.12006>
81. Stupinski AM, Alshaabi T, Arnold MV, Adams JL, Minot JR, Price M, Dodds PS, Danforth CM (2021) Quantifying language changes surrounding mental health on twitter. Available online at <https://arxiv.org/abs/2106.01481>
82. Minot JR, Cheney N, Maier M, Elbers D, Danforth CM, Dodds PS (2022) Interpretable bias mitigation for textual data: reducing gender bias in patient notes while maintaining classification performance. *ACM Trans Comput Healthc* 3:1–41. Available online at <https://arxiv.org/abs/2103.05841>
83. Ring JH IV, Van Oort CM, Durst S, White V, Near JP, Skalka C (2021) Methods for host-based intrusion detection with deep learning. *Digit Treats Res Pract* 2:1–29
84. Alshaabi T, Adams JL, Arnold MV, Minot JR, Dewhurst DR, Reagan AJ, Danforth CM, Dodds PS (2021) Storywrangler: a massive exploratorium for sociolinguistic, cultural, socioeconomic, and political timelines using Twitter. *Sci Adv* 7:eabe6534
85. Dodds PS, Minot JR, Arnold MV, Alshaabi T, Adams JL, Reagan AJ, Danforth CM (2020) Computational timeline reconstruction of the stories surrounding Trump: story turbulence, narrative control, and collective chronopathy. <https://arxiv.org/abs/2008.07301>
86. Dodds PS, Danforth CM (2009) Measuring the happiness of large-scale written expression: songs, blogs, and presidents. *J Happ Stud* 11(4):441–456
87. Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM (2011) Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. *PLoS ONE* 6:e26752
88. Dodds PS, Clark EM, Desu S, Frank MR, Reagan AJ, Williams JR, Mitchell L, Harris KD, Kloumann IM, Bagrow JP, Megerdooian K, McMahon MT, Tivnan BF, Danforth CM (2015) Human language reveals a universal positivity bias. *Proc Natl Acad Sci* 112(8):2389–2394. Available online at <http://www.pnas.org/content/112/8/2389>
89. Reagan AJ, Tivnan BF, Williams JR, Danforth CM, Dodds PS (2017) Sentiment analysis methods for understanding large-scale texts: a case for using continuum-scored words and word shift graphs. *EPJ Data Sci* 6:28
90. Gallagher RJ, Frank MR, Mitchell L, Schwartz AJ, Reagan AJ, Danforth CM, Dodds PS (2021) Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts. *EPJ Data Sci* 10:4. Available online at <https://arxiv.org/abs/2008.02250>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

Submit your next manuscript at ▶ [springeropen.com](https://www.springeropen.com)
