

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

Peter Sheridan Dodds,^{1,2,*} Joshua R. Minot,¹ Michael V. Arnold,¹ Thayer Alshaabi,¹ Jane Lydia Adams,¹ David Rushing Dewhurst,¹ Tyler J. Gray,^{1,2} Morgan R. Frank,³ Andrew J. Reagan,⁴ and Christopher M. Danforth^{1,2}

¹*Computational Story Lab, Vermont Complex Systems Center,
MassMutual Center of Excellence for Complex Systems and Data Science,
Vermont Advanced Computing Core, University of Vermont, Burlington, VT 05401.*

²*Department of Mathematics & Statistics, University of Vermont, Burlington, VT 05401.*

³*Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA, 02139*

⁴*MassMutual Data Science, Amherst, MA 01002.*

(Dated: February 25, 2020)

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. Comparisons of component size distributions for two complex systems—or a system with itself at two different time points—generally employ information-theoretic instruments, such as Jensen-Shannon divergence. We argue that these methods lack transparency and adjustability, and should not be applied when component probabilities are non-sensible or are problematic to estimate. Here, we introduce ‘allotaxonomy’ along with ‘rank-turbulence divergence’, a tunable instrument for comparing any two (Zipfian) ranked lists of components. We analytically develop our rank-based divergence in a series of steps, and then establish a rank-based allotaxonomograph 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 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.

I. INTRODUCTION

A. Instruments that capture complexity

Science stands on the ability to describe and explain, and precise quantification must ultimately secure any true understanding. 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.

But for complex phenomena made up of a great many types of components of greatly varying size—ecologies, stock markets, language—we must confront two major problems with our dashboards of simple instruments [1].

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 [2]. 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 [3], 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 [4].) 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 [5, 6].

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 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.

To contend with scale and internal diversity of complex systems, we need comprehensible, dynamically-adjusting dashboards. For comparisons of complex systems, we

* peter.dodds@uvm.edu

will argue for dynamic dashboards that have two core elements [7]:

1. A ‘big picture’ map-like overview; and
2. A ranking of components afforded by a tunable measure that is as plain-spoken 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; ‘allotaxometrics’ to refer to quantified allotaxonomy; and ‘allotaxonometers’ and ‘allotaxographs’ for the instruments of allotaxometrics.

B. Zipf 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 Zipf distributions (rather than laws), and what we will call ‘rank turbulence’—a phenomenon of system-system comparison. We describe and discuss these two common signatures of complex systems in turn.

In general, we will consider systems where each component type τ has at least one measurable—and hence rankable—“size” s_τ where size may be count, rate, physical size, monetary value, scoring in sports by individual players, and so on. 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 . 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 Sec. II A below.)

Zipf’s law is the specific observation that a Zipf ranking obeys a decaying power law [8–11]. That is, the size s_r of the r th ranked component obeys the scaling $s_r \sim r^{-\zeta}$ where the Zipf exponent is $\zeta > 0$. 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 Zipf distributions abound, with a few examples including word and phrase frequency in language [12, 13], city populations [8], node degrees in scale-free networks [14], firm size [15], and numbers of dependencies for software packages [16].

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 Zipf 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 [17, 18]; this is negated by a simple analysis [19],

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 [9, 16, 20, 21].

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 Zipf ranking s_r , there may well be a rich variation in the ordering of components [12, 22]. With this understanding, in earlier work by our group on comparing Zipf rankings of n -gram usage in large-scale texts, we introduced the concept of “lexical turbulence” [22]. 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 [23, 24]). 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’.

C. Motivation for a rank-based divergence

In comparing complex systems, why should we use component size ranks rather than probabilities or rates? Indeed, there is a smorgasbord of ways to compare two probability distributions for categorical data [25–27]. Ref. [26] catalogs around 60 probability-based comparisons which are variously distances, divergences, similarities, fidelities, and inner products. And Ref. [27] 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 [12, 28]. In natural ecological systems, for example, estimating the total number of organisms is famously difficult [28–31]. We can only then speak of relative rates and not absolute rates, and even then only for common enough species. For Twitter, subsampling 1-grams allows for robust estimation of the rates of common 1-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 [32–34].

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 Spearman’s rank correlation coefficient. All told, while in moving to rankings we may trade information for some simplification, we still preserve a great deal of meaningful structure.

Fifth and finally, rankings are an easily interpretable, ubiquitous construct. Ranked lists suffuse media surrounding entertainment (e.g., box office), music (Billboard charts), and sports.

The above notwithstanding, distances based on comparisons of Zipf rankings are to our knowledge relatively few, focus on traditional comparative metrics like Kendall’s Tau and Spearman’s rank correlation coefficient [35], and seem limited in application to extremely small systems, for example, comparing the top 20 to 50 ranked hits from two different search engines [35–37].

D. Paper outline

In Sec. II, we develop rank-turbulence divergence by (1) Establishing our notation and ranking process (Sec. II A); (2) Creating and explaining a specific kind of rank-rank histogram (Sec. II B); (3) Declaring a set of desired features for rank-turbulence divergence (Sec. II C); and then (4) Building and refining a rank-turbulence divergence that effectively captures these features (Sec. II D).

In Sec. III, 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 Flipbooks of allotaxonographs as supplementary online material on the arXiv and at 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 Sec. IV.

We present details of datasets and code in Sec. V, and

we round off our paper with some concluding thoughts in Sec. VI.

II. RANK-TURBULENCE DIVERGENCE

A. Notation, Ranking Methodology, and Exclusive Types

As mentioned in the introduction, we use Zipfian ranking [8], 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.

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 $D_\alpha^R(\Omega_1 \parallel \Omega_2)$. In Sec. II D, 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 their incorporation. Thus to help with clarity, 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:

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

We sort types by descending contribution, $\delta D_{\alpha,\tau}^R(R_1 \parallel R_2)$, indicating this ordering by the set $R_{1,2;\alpha}$.

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 [38]).

Therefore, when comparing two systems, we extend the list of types in both systems to be the union of the types for both. 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.

B. Rank-Rank Histograms for Basic Allotaxonomy

In Fig. 1A, we show an example of our base system-system comparison plot, what 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 Sec. V A for description of datasets).

To construct Fig. 1A, 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 simplified 1-grams observed on both days and thereby obtain rank-rank pairs for all 1-grams.

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 [40]), with the number of rank-rank pairs per cell increasing per the lower left scale in 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 [41]). The vertical and horizontal coordinates in the rotated histogram are proportional to $\log_{10} r_{\tau,1} r_{\tau,2}$ (measured downwards) and $\log_{10} r_{\tau,2}/r_{\tau,1}$ (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 with have 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 our present work, 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 or species abundance in two 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 choose one as a representative example.

To aid a user’s perception of what meaning might be rapidly conferred by a rank-rank histogram, we highlight a selection of the annotated words in Fig. 1A. 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.

Types appearing towards the top of the diamond rank high for both systems. For Fig. 1A, the 1-gram ‘RT’ is the most common word on both days: $r_{\text{RT},1} = r_{\text{RT},2} = 1$. Signifying retweet, ‘RT’ is an important—if Twitter-specific—functional structure, indicating the strength of echoing on Twitter. The words ‘the’ and ‘to’ are ranked 2nd and 3rd on both dates, while ‘and’ and ‘is’ are ranked 4th and 4th 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.

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 relative change in rank. On 2016/11/09, ‘Trump’ stands out relative to nearby words. Further down, ‘America’, ‘Donald’, ‘voted’, 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).

While 2016 election and Charlottesville terms dominate the sides of the histogram, unrelated names and events also appear. On the left we see the ‘gorilla’ (Harambe specifically) and ‘Meteorite’ while on the right, we find Lady Gaga and Zara Larsson (both performed concerts), and the Korean band BTS which was enjoying its rise to ultrafame over this time period [42].

The separated lines and points at the bottom of the histogram arise from logarithmic spacing. For systems with heavy-tailed Zipf distributions for discrete sizes, 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 [9]—the rare are legion.

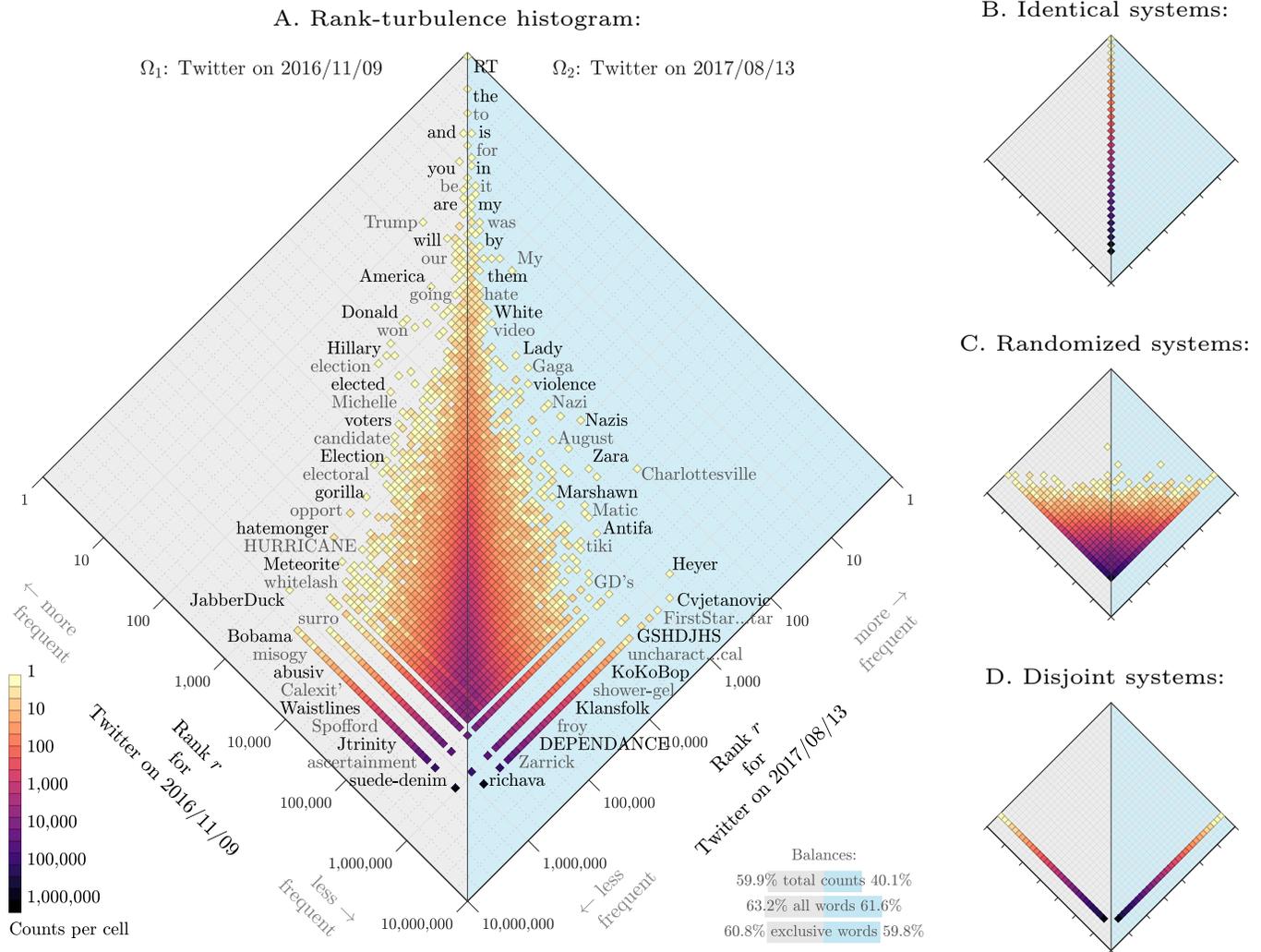


FIG. 1. **A.** An example allotaxonomic ‘rank-rank histogram’ comparing word usage ranks on two days of Twitter, 2016/11/09 and 2017/08/13. These dates are the day after the 2016 US presidential election and the day after the Charlottesville Unite the Right rally. Words are extracted first as 1-grams from tweets identified as English [39] and then filtered to match simple latin characters (see Sec. V A). 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 [40], and place a scale in the bottom left corner. 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 $r=9,149$ 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 distinct lines above the ‘exclusive-type lines’ correspond to words that appear once and twice in the other system. The three horizontal bars in the lower right show system balances. The top bar indicates the balance of total counts of words for each day: 59.9% versus 40.1%. The middle bar shows the percentage of the lexicon for the two days combined that appear on each day: 63.2% versus 61.6%. And the bottom bar shows the percentage of words on each day that are exclusive: 60.8% and 59.8%. **B–D.** The three rank-rank histograms on the right show the special, benchmark cases of: **B.** A Zipf ranking for compared with itself (vertical line; Ω_1); **C.** A ranked list versus a random shuffling of component types (Ω_1); and **D.** Two Zipf rankings for systems with no shared component types: a ‘vee’ structure (we 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**.

Moving upwards from the bottom, the three separated lines in Fig. 1A’s histogram correspond to words appearing zero times, once, and twice on the other side’s day. We define ‘exclusive types’ as those types that zero times in the other system, i.e., those types that appear along the bottom separated lines of the histogram.

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 ($r_{\text{Cvjetanovic},2}=672$). The word is the last name of a member of Identity Evropa who was part of the Unite the Right Rally; a photo of him holding a tiki torch and yelling was widely circulated [43]. 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 ($r_{\text{Cvjetanovic},1}=1,552,865$). 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.

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 in Fig. 1A, ‘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 (see the count scale in the lower left of 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 these of 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. 1B–D show examples of three extremes of how systems might compare on rank-rank histograms. For real-world data, we will see various imprints of these three limiting cases.

In Fig. 1B, we compare the Zipf ranking for identical systems (Ω_1 from Fig. 1A). The outcome is a colormap version of the system’s Zipf distribution arranged on the vertical $r_{\tau,1} = r_{\tau,2}$ line.

In Fig. 1C, 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. 1D, we compare Zipf 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 Zipf distributions forming a ‘vee’. We have already seen examples of these linear features in Fig. 1A. If system component lists are suffi-

ciently 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 Sec. III F, after introducing rank-turbulence divergence.

C. 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. 1A 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, and to 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: $D_{\alpha}^{\text{R}}(R_1 \parallel R_2) = D_{\alpha}^{\text{R}}(R_2 \parallel R_1)$.
3. Semi-positive: $D_{\alpha}^{\text{R}}(R_1 \parallel R_2) \geq 0$, and $D_{\alpha}^{\text{R}}(R_1 \parallel R_2) = 0$ 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 need 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. (1), each type τ additively contributes to rank-turbulence divergence a quantity $\delta D_{\alpha,\tau}^{\text{R}}(R_1 \parallel R_2)$, 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 $D_{\alpha}^{\text{R}}(\Omega_1 \parallel \Omega_2)$ 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. Zipfophilic: 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 [26, 27], in practice there are families of divergences on offer, and these have the potential to be adaptive and provide much more power and insight [27].
11. Storyfinding: Features 1–10 will ideally combine to help us rapidly see which types are most important in distinguishing two ranked lists.

D. 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 Zipfian rank is inversely related to its size. We thus will want to deal with inverses of ranks.

Given element τ has a Zipfian 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:

$$\left| \frac{1}{r_{\tau,1}} - \frac{1}{r_{\tau,2}} \right|. \quad (2)$$

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 α :

$$\left| \frac{1}{[r_{\tau,1}]^\alpha} - \frac{1}{[r_{\tau,2}]^\alpha} \right|^{1/\alpha}. \quad (3)$$

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 [44–46] and Hill numbers in ecology [6, 29].)

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. (3) does not yet behave as we might hope. We see that if $r_{\tau,1} \neq r_{\tau,2}$, Eq. (3) tends towards

$$\alpha^{1/\alpha} \left| \ln \frac{r_{\tau,1}}{r_{\tau,2}} \right|^{1/\alpha}, \quad (4)$$

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

In considering how to remedy this problematic limit, we observe that Eq. (4) contains a readily interpretable structure which we have already encountered in the preceding section: the log-ratio of ranks. In Sec. II B, we established a graphical interpretation for the rank-rank histogram in Fig. 1A. We identify $\left| \ln \frac{r_{\tau,1}}{r_{\tau,2}} \right| = |\ln r_{\tau,1} - \ln r_{\tau,2}|$ as being proportional to the horizontal distance from the $(\log_{10} r_{\tau,1}, \log_{10} r_{\tau,2})$ point to the vertical midline.

To preserve the core of Eq. (3), $\left| \frac{1}{[r_{\tau,1}]^\alpha} - \frac{1}{[r_{\tau,2}]^\alpha} \right|^{1/\alpha}$, maintain the form of the large α limit, fashion a well-behaved $\alpha \rightarrow 0$ limit, and to only use modifications that are monotonic in α , we introduce a prefactor and adjust the exponent in Eq. (3) as follows:

$$\frac{\alpha + 1}{\alpha} \left| \frac{1}{[r_{\tau,1}]^\alpha} - \frac{1}{[r_{\tau,2}]^\alpha} \right|^{1/(\alpha+1)}. \quad (5)$$

The $\alpha \rightarrow 0$ limit is now simply $\left| \ln \frac{r_{\tau,1}}{r_{\tau,2}} \right|$, while the $\alpha \rightarrow \infty$ limit is unchanged. (We note that an alternate modification of simply introducing a prefactor of $\alpha^{-1/\alpha}$ to Eq. (3) 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:

$$\begin{aligned} D_\alpha^R(R_1 \| R_2) &= \sum_{\tau \in R_{1,2;\alpha}} \delta D_{\alpha,\tau}^R(R_1 \| R_2) \\ &= \frac{1}{\mathcal{N}_{1,2;\alpha}} \frac{\alpha + 1}{\alpha} \sum_{\tau \in R_{1,2;\alpha}} \left| \frac{1}{[r_{\tau,1}]^\alpha} - \frac{1}{[r_{\tau,2}]^\alpha} \right|^{1/(\alpha+1)}. \end{aligned} \quad (6)$$

While analytic forms for the normalization factor $\mathcal{N}_{1,2;\alpha}$ could be constructed, we take a numerical approach. We compute $\mathcal{N}_{1,2;\alpha}$ by taking the two systems to be disjoint while maintaining their underlying Zipf distributions. 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 Zipf distributions.

To determine $\mathcal{N}_{1,2;\alpha}$, we observe that if the Zipf 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

is then:

$$\begin{aligned} \mathcal{N}_{1,2;\alpha} &= \frac{\alpha+1}{\alpha} \sum_{\tau \in R_1} \left| \frac{1}{[r_{\tau,1}]^\alpha} - \frac{1}{[N_1 + \frac{1}{2}N_2]^\alpha} \right|^{1/(\alpha+1)} \\ &+ \frac{\alpha+1}{\alpha} \sum_{\tau \in R_1} \left| \frac{1}{[N_2 + \frac{1}{2}N_1]^\alpha} - \frac{1}{[r_{\tau,2}]^\alpha} \right|^{1/(\alpha+1)}. \end{aligned} \quad (7)$$

We note that for a disjoint pair of systems, their randomized versions will necessarily still be disjoint, and $D_{\alpha;\text{rand}}^{\text{R}}(R_1 \| R_2) = 1$.

E. 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. (6), in the limit of $\alpha \rightarrow 0$, we have

$$D_0^{\text{R}}(R_1 \| R_2) = \sum_{\tau \in R_{1,2;\alpha}} \delta D_{0,\tau}^{\text{R}} = \frac{1}{\mathcal{N}_{1,2;0}} \sum_{\tau \in R_{1,2;\alpha}} \left| \ln \frac{r_{\tau,1}}{r_{\tau,2}} \right|, \quad (8)$$

where

$$\mathcal{N}_{1,2;0} = \sum_{\tau \in R_1} \left| \ln \frac{r_{\tau,1}}{N_1 + \frac{1}{2}N_2} \right| + \sum_{\tau \in R_2} \left| \ln \frac{r_{\tau,2}}{\frac{1}{2}N_1 + N_2} \right|. \quad (9)$$

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 [25].

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

$$\begin{aligned} D_\infty^{\text{R}}(R_1 \| R_2) &= \sum_{\tau \in R_{1,2;\alpha}} \delta D_{\infty,\tau}^{\text{R}} \\ &= \frac{1}{\mathcal{N}_{1,2;\infty}} \sum_{\tau \in R_{1,2;\alpha}} (1 - \delta_{r_{\tau,1}r_{\tau,2}}) \max \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\}. \end{aligned} \quad (10)$$

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

$$\mathcal{N}_{1,2;\infty} = \sum_{\tau \in R_1} \frac{1}{r_{\tau,1}} + \sum_{\tau \in R_2} \frac{1}{r_{\tau,2}}. \quad (11)$$

For probability-based divergences, the $\alpha = \infty$ limit for rank-turbulence divergence aligns with the Motyka distance [25, 26].

Because we are interested in real, finite systems, we are not concerned with convergence. Nevertheless, with appropriate treatment, infinite theoretical systems could be evaluated.

III. RANK-TURBULENCE DIVERGENCE GRAPHS AS ALLOTAXONOMETRIC INSTRUMENTS

A. Anatomy of an allotaxonograph with word usage on Twitter as an example

We now combine rank-rank histograms with rank-turbulence divergence to generate a tunable single-parameter instrument for exploring how two systems differ. In Fig. 2, we present a ‘rank-turbulence divergence graph’ as an example allotaxonograph. We again compare the two days of Twitter—the 2016 US election with the 2017 Charlottesville riots—that we examined in Sec.II B.

There are two main components to our general divergence graphs: A map-like histogram and an ordered list of types contributing the most to the divergence measure being explored.

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. In all our divergence graphs, we include the divergence's expression above the top left of the histogram. We overlay the histogram with contour lines of constant $\delta D_{1/3,\tau}^{\text{R}}$. The contour lines are chosen so that they are evenly spaced and anchored along the bottom two axes, making for simple tracking as α is varied. The inset to the upper right of the histogram provides a scale for values of $\delta D_{1/3,\tau}^{\text{R}}$.

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 the top left Fig. 2 shows 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.

As they are independent of divergence measures, the annotations and their locations on the histogram remain unchanged from Fig. 1. We now incorporate a linear gray scale based on $\delta D_{1/3,\tau}^{\text{R}}$, with higher scoring words accentuated, lower scoring words faded. We now see ‘Trump’ and ‘Charlottesville’ stand out. 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 in Fig. 2. We order the top 40 words by decreasing value of $\delta D_{1/3,\tau}^{\text{R}}$, indicated by the underlying

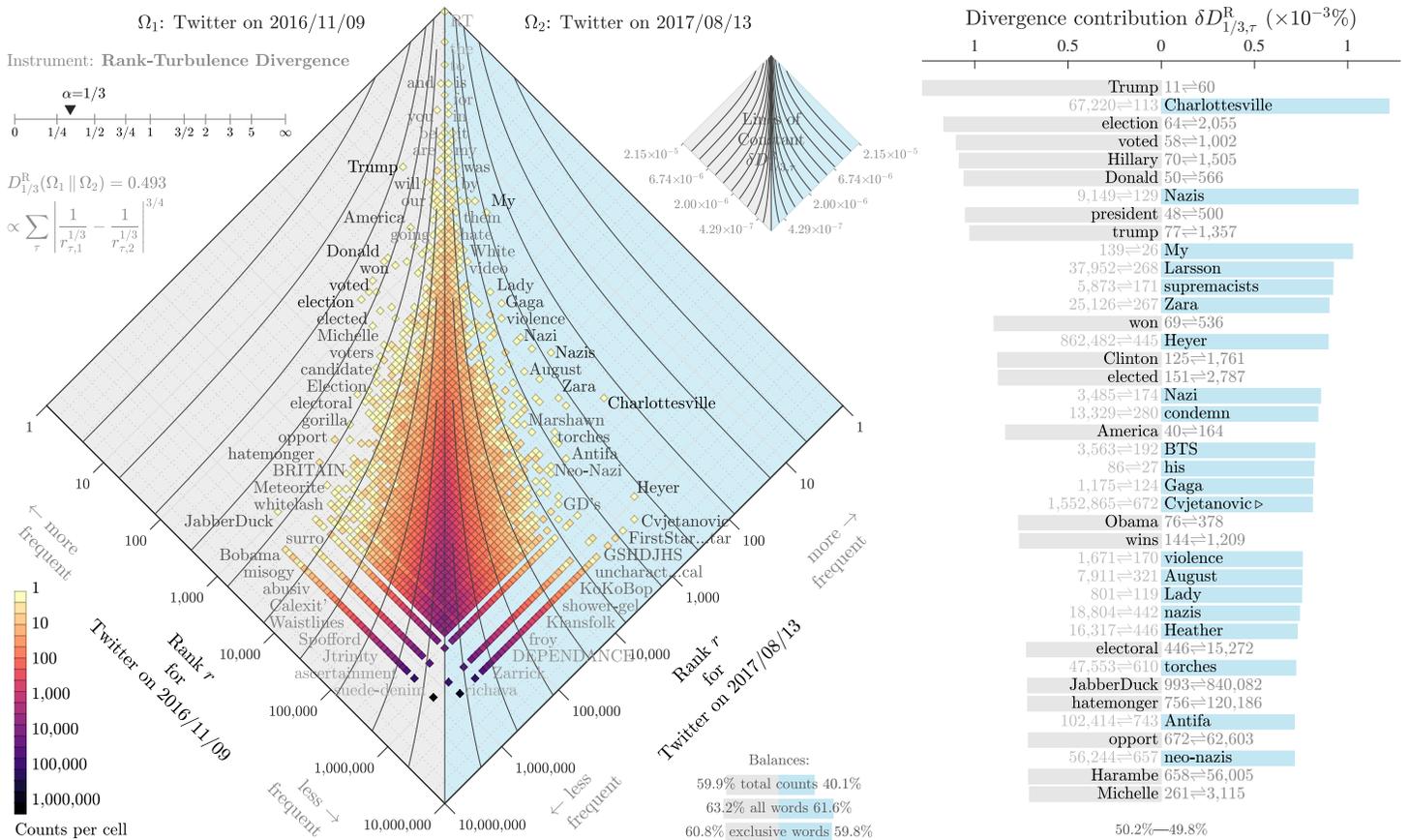


FIG. 2. Example allotaxonograph using rank-turbulence divergence to compare word usage on different days of Twitter. We examine the same dates of the 2016 US Presidential Election and the Charlottesville Unite the Right rally as the rank-rank histogram of Fig. 1. We add rank-turbulence divergence to Fig. 1’s histogram with an overlay of contour lines, a gauge for α and the expression for $D_{1/3}^R$ in the upper left corner, and a scale for the contour lines in the upper right. Based on contributions of each word to $D_{1/3}^R$, we generate the ordered list on the right by descending values of $\delta D_{1/3,\tau}^R$. Words are arranged left and right and colored gray and blue in accordance with the date on which they are most prevalent. The two dates’ ranks for each word in the list are indicated on the opposite side. For example, $r_{\text{Trump},1}=11$ and $r_{\text{Trump},2}=60$, and $r_{\text{Heyer},1}=862,482$ and $r_{\text{Heyer},2}=445$. While an exact match is intended, a few annotated words on the histogram differ from Fig. 1 due to chance (e.g., ‘HURRICANE’ and ‘BRITAIN’ on the left side). The instrument’s function and layout are highly configurable in our figure-building script. For example, the choice of divergence (rank or otherwise), axis limits, maximum length of type names, histogram cell size, and the guide adornments ‘less talked about’ and ‘more talked about’ are all system-specific settings. As a design choice, we limit the resolution of α to multiples of $1/12$. For further details on the underlying histogram, see the caption of Fig. 1.

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 [42]), 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.

For systems for which we are confident we have determined the constituent elements, 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. In the ordered list, we indicate types that appear in only of the two systems 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’. 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 ∞ .

In all allotaxonographs, we show three kinds of balances at the bottom right of the rank-rank histogram.

First, we see the breakdown of total counts between the two dates at 59.9% and 40.1% (the election generated more tweets than Charlottesville). Second, we have that all words in the lexicon for the two days combined, just over 60% appear on each of the two days. Third, we create separate lexicons for each day, and find that around 60% are exclusive for both days, giving a sense of strong turnover. As we will see, these balances can vary greatly across system comparisons.

B. 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 α values 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 though 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.

C. 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 [52] for five-year censuses completed in 1985 and 2015 ($\Omega^{(1)}$ and $\Omega^{(2)}$) [47].

In being visually close to the limit of comparing two identical rankings (Fig. 1B), 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 systems [53], 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 compare the 1985 and 2015 distributions by applying rank-turbulence divergence with $\alpha = 0$. The overall score $D_0^R(R_1 \| R_2) = 0.077$ is well short of the randomized equivalent of $D_{0;\text{rand}}^R(R_1 \| R_2) = 0.376$ (from 100 samples; standard deviation $\sigma=0.012$). Per Eq. (8), 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* [48–51]—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.

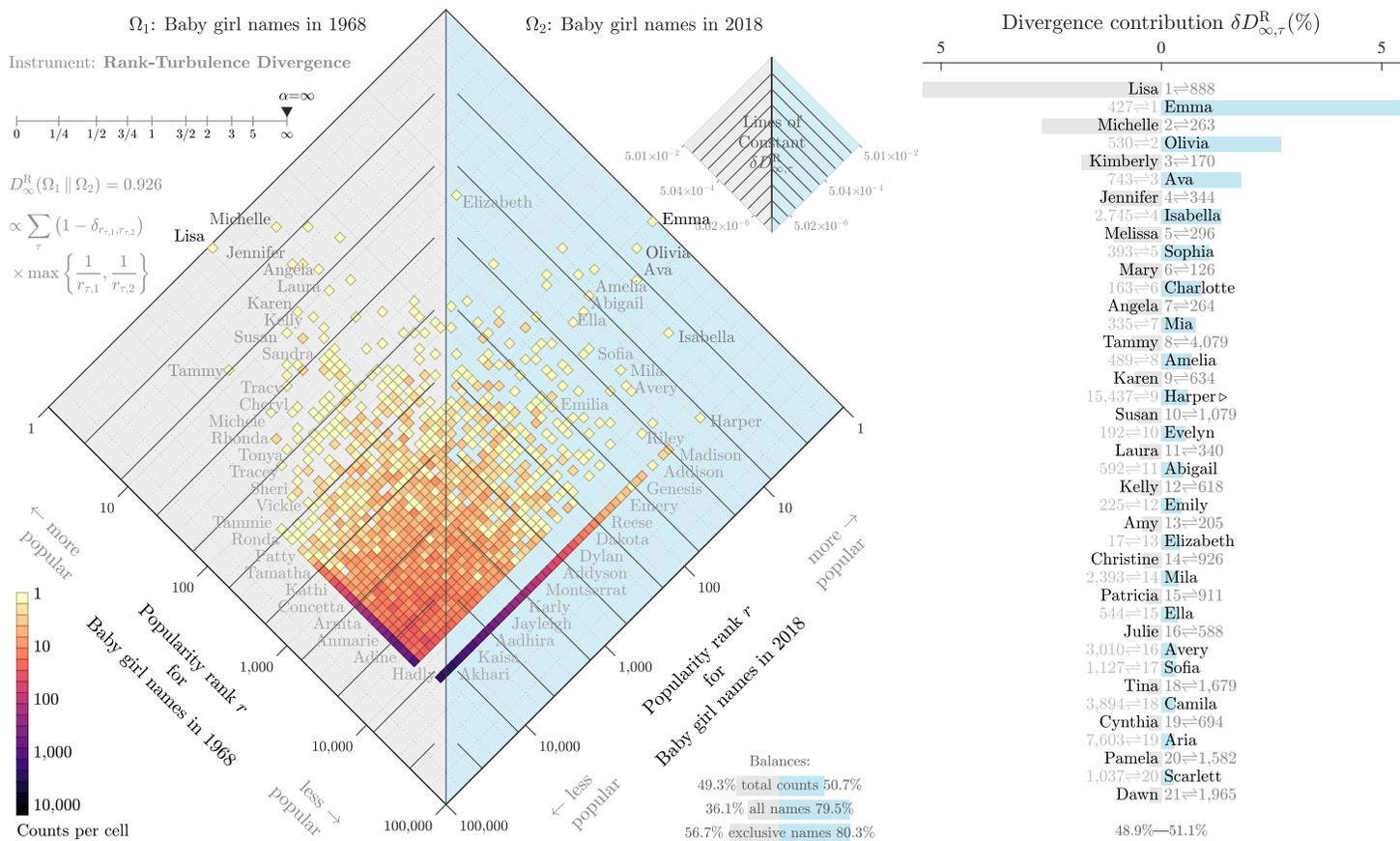


FIG. 4. **Allotaxonograph comparing names of girls born in the US in 1968 and 2018.** For dataset details, see Sec. V A. Of our four main case studies, baby name distributions show the strongest change with $D_\alpha^R(\Omega_1 \parallel \Omega_2)$ scores verging on that of the random equivalent. The 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, ‘Hadly’ is a 2018 exclusive word but it is oriented towards the left per our annotation method (see Fig. 1 and Sec. II B). 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.

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

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. Overall, the turnover is stronger for girl names than boy names. We can get 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. The balance quantities show one major difference: For girls, 56.7% of 1968 names are exclusive to 1968 while for boys, the same quantity is only 36.5%.

For girls, ranging from common 2018 names (‘Harper’, ‘Madison’, and ‘Addison’) down to rare names (‘Kaisa’, ‘Akhari’, and ‘Hadly’), the 2018 exclusive names comprise 80.3% of all names (14,485 of 18,029). For the smaller name base of boys, we see 10,994 of 14,004 (78.5%) 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 (4,650 of 8,194) and 36.5% of all boy

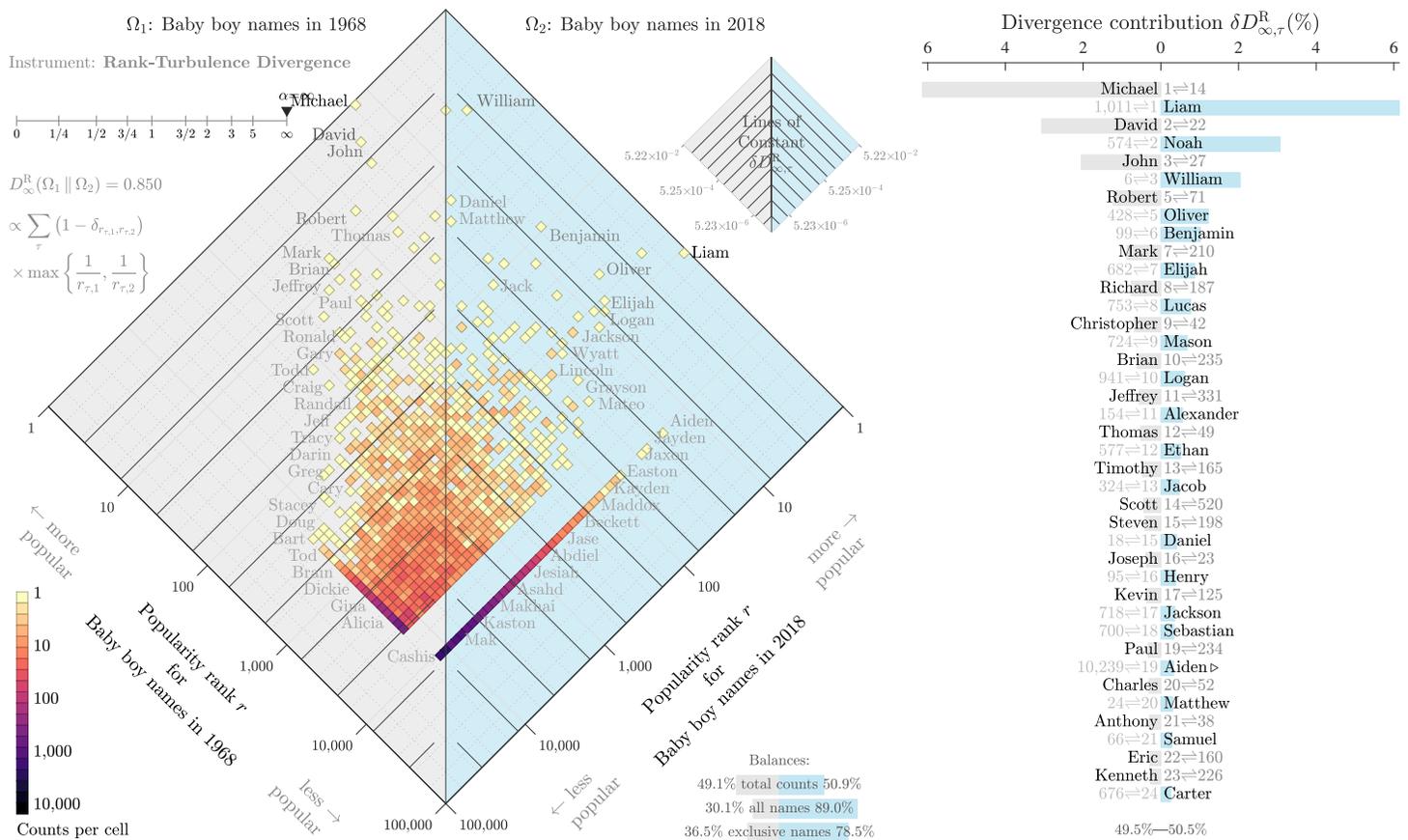


FIG. 5. Allotaxonograph comparing US boy names for the years 1968 and 2018. For dataset details, see Sec. V A. At the bottom of the histogram, ‘Cashis’ is oriented to the left but is a 2018 exclusive word, is per ‘Hadly’ in Fig. 4. As for girl names, we provide two Flipbooks showing 50 year gap comparisons moving through time (Flipbook S6) and the effects of varying α for the 1968–2018 comparison (Flipbook S8).

names (1,732 of 4,742) 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 note that the asymmetries of both histograms—their apparent right-side ‘heaviness’—are not due even in part to changes in overall numbers. 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 unique names in the years 1968 and 2018 are strikingly different however: 8,194 and 18,029 for girls (120% increase), and 4,742 and 14,004 for boys (195% increase). Two of the major factors which lead to this explosion in name-space are immigration and creation of new names.

Using the overall birth numbers, we can estimate the percentage of names absent from our dataset—those with

less than 5 instances: 4.06% for 1968 and 8.62% for 2018 for girls, and 2.11% for 1968 and 6.66% for 2018 for boys. The 2018 Zipf 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 $D_{\infty}^R(R_1 \parallel R_2) = 0.926$. For the same time frame comparison, boy names have a lesser but still high value of $D_{\infty}^R(R_1 \parallel R_2) = 0.850$. Both values are below but not far from the randomized equivalents with Zipf distributions held constant: $D_{\infty; \text{rand}}^R(R_1 \parallel R_2) = 0.973$ 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 Zipf 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 maximum rank 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 through to 1961.

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 make 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’.

Finally, our allotaxonomic instrument 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 $/[Aa][iy]*d+[aeiouy]n+$/$. A selection of examples ranging from common to rare, highlighting variations on Brayden, are: ‘Aiden’ ($r=19$) ‘Jayden’ (30), ‘Brayden’ (84), ‘Kayden’ (97), ‘Zayden’ (185.5), ‘Rayden’ (683), ‘Braydon’ (856), ‘Braidon’ (1,239), ‘Bradyn’ (1,936), ‘Grayden’ (1,936), ‘Braydan’ (3,534.5), ‘Braydin’ (3,817.5), ‘Bladen’ (4,974.5), ‘Blayden’ (5,177), ‘Braidyn’ (5,177), ‘Vayden’ (5,870), ‘Braydyn’ (6,873), ‘Wayden’ (7,322), ‘Bradon’ (8,434.5), ‘Slayden’ (8,434.5), ‘Xzayden’ (10,155.5), Blaidon’ (11,389.5), ‘Braydenn’ (13,042), and ‘Braidon’ (13042).

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=1,152.5$), ‘Jacalyn’ (1,528.5), and ‘Cherilyn’ (1,870.5), 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 and syllables.

E. Allotaxonomy of publicly traded US companies: Stability, shocks, and errors

In Fig. 6, 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 Sec. V A). The allotaxonomograph 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 allotaxonomographs for market cap comparisons for 6 year time gaps starting 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 Sec. III C, 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 Mobil ($1 \rightarrow 9$), and AT&T ($4 \rightarrow 19$). Berkshire Hathaway’s apparent drop stems from a dataset error which we discuss below. On the right side

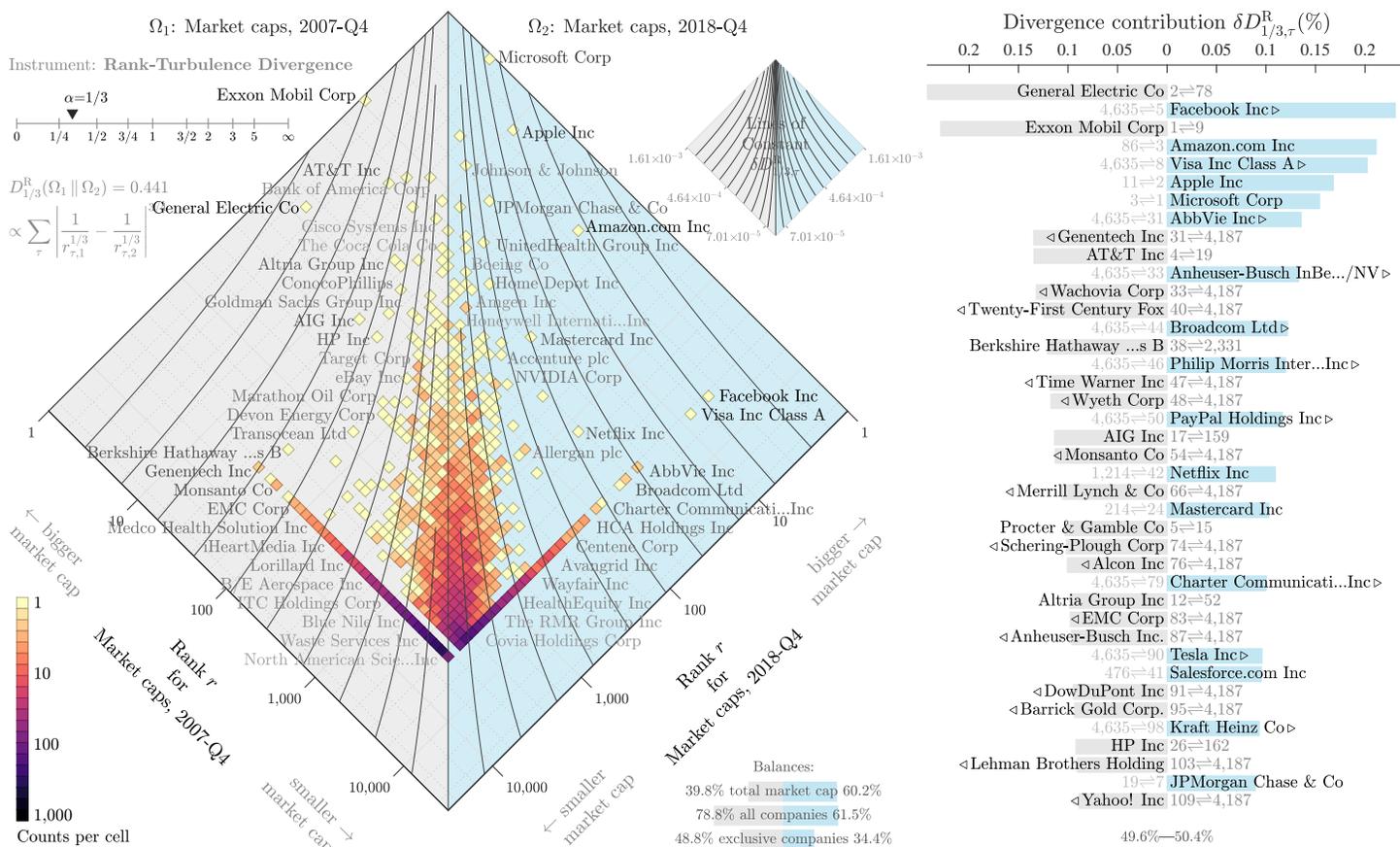


FIG. 6. **Allotaxonomic comparison of publicly traded US companies in 2007 and 2018 by fourth quarter market capitalization.** The rank-rank histogram is a hybrid of a vertical structure we see for relatively stable systems (Fig. 1B), and a ‘vee’ of disjoint systems (Fig. 1D). The disjoint feature results from sharp transitions as companies fail, merge with or are acquired by others, or go public or return to private, but also from missing and misrecorded data. Berkshire Hathaway’s market cap, for example, was misrecorded as a thousand fold drop. We include Berkshire Hathaway and other errors in part to show how an allotaxonomic analysis can sharply reveal dataset problems. See Sec. III F for discussion, and Sec. V A for dataset details.

for companies in existence in both 2007 and 2018, technology companies dominate: Amazon ($r=86 \rightarrow 3$), Apple ($11 \rightarrow 2$), Microsoft ($3 \rightarrow 1$), and Netflix ($1,214 \rightarrow 42$).

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 stand-

out entrants. With respective initial public offerings (IPOs) in 2008 and 2012, we find them 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 $\alpha = 1/3$ ($r=1,214 \rightarrow 24$). 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 \$528,336.12M to \$335.04M at the end of 2018).

We have chosen to leave such errors in Fig. 6 to help demonstrate the importance of using a rich, graphical allotaxonomic instrument. With a naive measurement

of divergence, we would easily miss problematic data points. Evidently, further cleaning of the market cap dataset would be required for further investigations.

The market cap histogram shows the importance of using a rich allotaxonomic instrument, and how we must take care when measuring divergences of any kind. The histogram’s form is not as simple as those we have seen for Twitter, species abundance, and baby names, and it would be evidently problematic to allow for an unexamined, automated fitting of α for rank-turbulence divergence (or parameters of any other divergence).

Given the composite form of the allotaxonograph 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 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.

F. Truncation Effects for Rank-Based Allotaxonographs

Truncation of a system’s Zipf distributions is a common if often overlooked problem [28, 56]. 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 [9]. The Google Books n -gram corpus only includes n -grams which have appeared 40 or more times [57], 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 put together a subset of the rank-rank histograms to form Fig. 7.

The five rows of Fig. 7 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 random structure for baby names and the stable vertical structure for market caps start to be revealed by $N=1,000$.

For the Flipbooks, we use the same values of α for Twitter $\alpha=1/3$, tree species $\alpha=0$, and market caps $\alpha=1/3$ (Figs. 2, 3, and 6). For baby names, we take the $\alpha = 0$ limit as this is the most challenging for the truncated version.

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.

IV. 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 best ‘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 (Sec. III F).

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%; $\alpha = 1/3$.

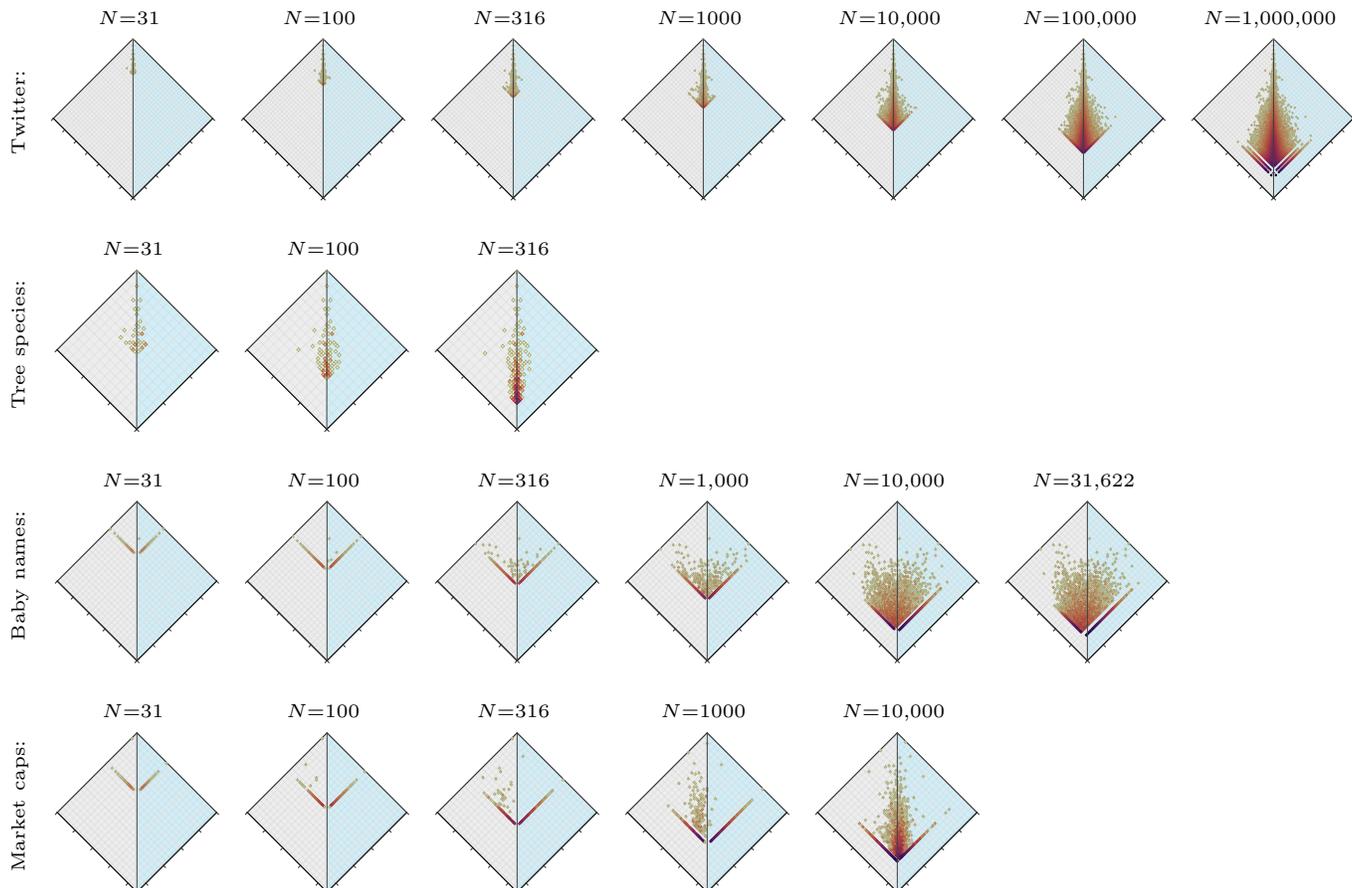


FIG. 7. **Exploration of the effect of subsampling data for allotaxonomic analyses.** The rows correspond to the four case studies of Twitter, trees, baby names, and market caps (see Figs. 2–6). Each row shows abstracted rank-rank histograms for Zipf distribution truncations to the top N types. As N increases, the Twitter and tree species histograms are revealed in a clean fashion, while baby names and market caps begin with a disjoint system ‘vee’ that masks their large N forms. Rows extend to the maximum system size for each comparison, and all colormaps and limits correspond to those used for the four case studies. For allotaxonomic analyses, see Flipbooks S10–S14.

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 Sec. IIID, comparisons of baby girl name distributions 50 years apart starting in 1880 and going forward in 5 year increments, with $\alpha = 1/3$. Ends with Fig. 4.

Flipbook S7—Baby boy names over time: Described in Sec. IIID, comparisons of baby girl name distributions 50 years apart starting in 1880 and going forward in 5 year increments, with $\alpha = 1/3$. Ends with Fig. 5.

Flipbook S6—Baby girl names, 1968–2018:

Described in Sec. IIID, shows effect of varying α , with Fig. 4 as the fifth page.

Flipbook S8—Baby boy names, 1968–2018: Described in Sec. IIID, 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. 7 with $\alpha = 1/3$.

Flipbook S11—Tree species abundance, truncated: Full series of allotaxonographs corresponding to histograms of row 2 in Fig. 7 with $\alpha = 0$.

Flipbook S12—Baby girl names, truncated: Full series of allotaxonographs corresponding to histograms of row 3 in Fig. 7 with $\alpha = \infty$.

Flipbook S13—Baby boy names, truncated: Full

series of allotaxonographs corresponding to histograms of row 4 in Fig. 7 with $\alpha = \infty$.

Flipbook S14—Market caps, truncated: Full series of allotaxonographs corresponding to histograms of row 5 in Fig. 7 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 $\delta D_{1/3,7}^R$, 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. [58], rendering the Google Books project unreliable, as is. Nevertheless, the Version 2 n -grams dataset for English fiction is worth exploring [22] 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 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 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 allotaxonometry.

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 $\alpha = 1/3$).

V. DATA AND CODE

A. 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. [39].

Species abundance on Barro Colorado Island: The dataset and its online repository for censuses taken over 35 years are described in Ref. [47].

Baby names: Data taken from Social Security Card applications. For each year from 1880–2018, 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://ssa.gov/oact/babynames/numberUSbirths.html>.

Market cap data: The underlying dataset comprises 9,322 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 [57]: <https://books.google.com/ngrams>.

Causes of Death in Hong Kong The dataset is described in Ref. [59–61] and has been well studied by others [62–67]. 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.

B. 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 Version R2019b. 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 [68], and generalized symmetric entropy divergence which includes Jensen-Shannon divergence as a special case.

VI. CONCLUDING REMARKS

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 Sec. IC, 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 allotaxonographs built around our conception of a tunable rank-turbulence divergence. In Tab. I, 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. (6) is the interpretable difference of inverse powers of type ranks:

$$\left| \frac{1}{[r_{\tau,1}]^\alpha} - \frac{1}{[r_{\tau,2}]^\alpha} \right|. \quad (12)$$

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 [26] and imposing the Zipfian 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 automatically 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 [7].

Per our introduction, there is tendency across diverse fields towards creating single-number measurements of complex systems, and that this is especially problematic when heavy-tailed Zipf distributions are in evidence. 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 all provided rich and diverse examples of allotaxonomic comparisons. Our ability

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

TABLE I. A selection of example system comparisons producing a range of $D_\alpha^R(\Omega_1 \parallel \Omega_2)$ values.

to readily analyze the effects of partially sampled data in Sec. III F further showed the value of a rank-based approach.

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.

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.

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 and other probability-based divergences including the Jensen-Shannon divergence and its generalizations in [68].

When rank turbulence is in evidence, as in the case

of Twitter, 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 [6, 29, 45, 46], we have found that for system comparisons, the optimal value of α , if it exists, is dependent on the pair of systems being compared.

In the present work, we have left open the possibility of an analytic connection between the rank-turbulence scaling described at the end of Sec. I B, and, to the extent that well-defined scaling is present, with an optimal α for rank-turbulence divergence.

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

$$p^\alpha \sim 1/r^\alpha = \exp\{-\alpha E/T\} = \exp\{-E/T'\}, \quad (13)$$

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.

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.

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 systems 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.

ACKNOWLEDGMENTS

The authors are grateful for support furnished by MassMutual and Google, and the computational facilities provided by the Vermont Advanced Computing Core. PSD thanks J. S. Weitz for unbidden and bidden abuse in the tradition of K. X. Whipple.

-
- [1] D. Borland, W. Wang, J. Wang, J. Shrestha, and D. Gotz, “Selection bias tracking and detailed subset comparison for high-dimensional data,” (2019), available online at <https://arxiv.org/abs/1906.07625>.
- [2] J. M. Diamond, *Guns, Germs, and Steel* (W. W. Norton & Company, 1997).
- [3] P. Turchin, T. E. Currie, H. Whitehouse, P. François, K. Feeney, D. Mullins, D. Hoyer, C. Collins, S. Grohmann, P. Savage, *et al.*, Proceedings of the National Academy of Sciences **115**, E144 (2018).
- [4] G. Strang, *Introduction to Linear Algebra*, 4th ed. (Cambridge Wellesley Press, Wellesley, MA, 2009).
- [5] C. E. Shannon, The Bell System Tech. J. **27**, 379 (1948).
- [6] L. Jost, *Oikos* **113**, 363 (2006).
- [7] S. E. Alajajian, J. R. Williams, A. J. Reagan, S. C. Alajajian, M. R. Frank, L. Mitchell, J. Lahne, C. M. Danforth, and P. S. Dodds, PLoS ONE **12**, e0168893 (2017), arXiv version available at <http://arxiv.org/abs/1507.05098>.
- [8] G. K. Zipf, *Human Behaviour and the Principle of Least-Effort*, patterns (Addison-Wesley, Cambridge, MA, 1949).
- [9] H. A. Simon, *Biometrika* **42**, 425 (1955).
- [10] M. E. J. Newman, *Contemporary Physics* **46**, 323 (2005).
- [11] B. Coromina-Murtra and R. Solé, *Physical Review E* **82**, 011102 (2010).
- [12] M. Gerlach, F. Font-Clos, and E. G. Altmann, *Physical Review X* **6**, 021009 (2016).
- [13] J. R. Williams, P. R. Lessard, S. Desu, E. M. Clark, J. P. Bagrow, C. M. Danforth, and P. S. Dodds, *Nature Scientific Reports* **5**, 12209 (2015).
- [14] A.-L. Barabási and R. Albert, *Science* **286**, 509 (1999).
- [15] R. Axtell, *Science* **293**, 1818 (2001).
- [16] T. Maillart, D. Sornette, S. Spaeth, and G. von Krogh, *Phys. Rev. Lett.* **101**, 218701 (2008).
- [17] G. A. Miller, *American Journal of Psychology* **70**, 311 (1957).
- [18] G. A. Miller, “Introduction to reprint of G. K. Zipf’s ‘The Psycho-Biology of Language.’ MIT Press, Cambridge MA,” (1965).
- [19] R. Ferrer-i-Cancho and B. Elvevåg, PLoS ONE **5**, e9411 (2010).
- [20] B. B. Mandelbrot, in *Communication Theory*, edited by W. Jackson (Butterworth, Woburn, MA, 1953) pp. 486–502.
- [21] P. S. Dodds, D. R. Dewhurst, F. F. Hazlehurst, C. M. Van Oort, L. Mitchell, A. J. Reagan, J. R. Williams, and C. M. Danforth, *Physical Review E* **95**, 052301 (2017).
- [22] E. A. Pechenick, C. M. Danforth, and P. S. Dodds, *Journal of Computational Science* **21**, 24 (2017), available online at <http://arxiv.org/abs/1503.03512>.
- [23] R. Ferrer-i-Cancho and R. V. Solé, *Journal of Quantitative Linguistics* **8**, 165 (2001).
- [24] J. R. Williams, J. P. Bagrow, C. M. Danforth, and P. S. Dodds, *Physical Review E* **91**, 052811 (2015).
- [25] M.-M. Deza and E. Deza, *Dictionary of distances* (Elsevier, 2006).
- [26] S.-H. Cha, *International Journal of Mathematical Models and Methods in Applied Sciences* **1**, 300 (2007).
- [27] A. Cichocki and S.-i. Amari, *Entropy* **12**, 1532 (2010).
- [28] B. Haegeman, J. Hamelin, J. Moriarty, P. Neal, J. Dushoff, and J. S. Weitz, *The ISME journal* **7**, 1092 (2013).
- [29] M. O. Hill, *Ecology* **54**, 427 (1973).
- [30] N. J. Gotelli and R. K. Colwell, *Biological diversity: Frontiers in measurement and assessment* **12**, 39 (2011).
- [31] A. Chao, N. J. Gotelli, T. Hsieh, E. L. Sander, K. Ma, R. K. Colwell, and A. M. Ellison, *Ecological monographs* **84**, 45 (2014).
- [32] S. Merritt and A. Clauset, *EPJ Data Science* **3** (2014).
- [33] A. Clauset, M. Kogan, and S. Redner, *Phys. Rev. E* **91**, 062815 (2015).
- [34] D. P. Kiley, A. J. Reagan, L. Mitchell, C. M. Danforth, and P. S. Dodds, *Physical Review E* **93** (2016), available online at <http://arxiv.org/abs/1507.03886>.
- [35] R. Fagin, R. Kumar, and D. Sivakumar, *SIAM Journal on discrete mathematics* **17**, 134 (2003).
- [36] J. Bar-Ilan, M. Mat-Hassan, and M. Levene, *Computer networks* **50**, 1448 (2006).
- [37] W. Webber, A. Moffat, and J. Zobel, *ACM Transactions on Information Systems (TOIS)* **28**, 1 (2010).
- [38] T. J. Gray, C. M. Danforth, and P. S. Dodds, “Hahahahaha, Duuuuude, Yeeessss!: A two-parameter characterization of stretchable words and the dynamics of mistypings and misspellings,” (2019), available online at <https://arxiv.org/abs/1907.03920>.
- [39] T. Alshaabi *et al.*, “Temporal dynamics of 170 languages on twitter from 2008–2020,” (2020), forthcoming.
- [40] Y. Liu and J. Heer, in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (ACM, 2018) p. 598.
- [41] C. T. Bergstrom and J. D. West, “Why scatter plots suggest causality, and what we can do about it,” (2018), available online at <https://arxiv.org/abs/1809.09328>.
- [42] P. S. Dodds, J. R. Minot, M. V. Arnold, T. Alshaabi, J. L. Adams, D. R. Dewhurst, A. J. Reagan, and C. M. Danforth, “Fame and Ultrafame: Measuring and comparing daily levels of ‘being talked about’ for United States’ presidents, their rivals, God,

- countries, and K-pop,” (2019), available online at <https://arxiv.org/abs/1910.00149>.
- [43] “Identity Evropa,” Wikipedia (2019), https://en.wikipedia.org/w/index.php?title=Identity_Evropa&oldid=934670726; Accessed on 2020/01/28.
- [44] A. Rényi, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* (1961).
- [45] C. Tsallis, in *Nonextensive statistical mechanics and its applications* (Springer, 2001) pp. 3–98.
- [46] C. Keylock, *Oikos* **109**, 203 (2005).
- [47] R. Condit *et al.*, “Complete data from the barro colorado 50-ha plot: 423617 trees, 35 years, v3, dataone dash, dataset,” (2019).
- [48] W. Trelease, *Systematic plant studies: mainly tropical America* (1927).
- [49] P. C. Standley, *Smithsonian Miscellaneous Collections* (1927).
- [50] W. Thies and E. K. V. Kalko, *Oikos* **104**, 362 (2004).
- [51] T. Y. Andrade, W. Thies, P. K. Rogeri, E. K. V. Kalko, and M. A. R. Mello, *Journal of Mammalogy* **94**, 1094 (2013).
- [52] R. Condit, P. Ashton, P. Baker, S. Sarayudh B., Gunatilleke, N. Gunatilleke, S. Hubbell, R. Foster, A. Itoh, J. LaFrankie, H. Lee, E. Losos, N. Manokaran, R. Sukumar, and T. Yamakura., *Science* **288**, 1414 (2000).
- [53] S. H. Strogatz, *Nonlinear Dynamics and Chaos*, general (Addison Wesley, Reading, Massachusetts, 1994).
- [54] M. W. Hahn and R. A. Bentley, *Proceedings of the Royal Society of London. Series B: Biological Sciences* **270**, S120 (2003).
- [55] M. Wattenberg, in *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005*. (IEEE, 2005) pp. 1–7.
- [56] A. Koplenig, S. Wolfer, and C. Müller-Spitzer, *Entropy* **21**, 464 (2019).
- [57] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. A. Lieberman, *Science Magazine* **331**, 176 (2011).
- [58] E. A. Pechenick, C. M. Danforth, and P. S. Dodds, *PLoS ONE* **10**, e0137041 (2015).
- [59] “Micro-data set of known and registered deaths,” https://www.censtatd.gov.hk/service_desk/list/microdata/index.jsp (2018), data retrieved from Census and Statistics Department of Hong Kong.
- [60] “District and constituency area,” <https://www.bycensus2016.gov.hk/en/bc-dp.html> (2016), data retrieved from Census and Statistics Department of Hong Kong.
- [61] “Tertiary planning units,” <https://www.bycensus2016.gov.hk/en/bc-dp-tpu.html> (2016), data retrieved from Census and Statistics Department of Hong Kong.
- [62] C.-M. Wong, S. Ma, A. J. Hedley, and T.-H. Lam, *Environmental health perspectives* **109**, 335 (2001).
- [63] T.-H. Lam, S.-Y. Ho, A. J. Hedley, K.-H. Mak, and G. M. Leung, *Annals of epidemiology* **14**, 391 (2004).
- [64] C.-Q. Ou, A. J. Hedley, R. Y. Chung, T.-Q. Thach, Y.-K. Chau, K.-P. Chan, L. Yang, S.-Y. Ho, C.-M. Wong, and T.-H. Lam, *Environmental research* **107**, 237 (2008).
- [65] H. Qiu, L. Tian, K.-f. Ho, V. C. Pun, X. Wang, and T. Ignatius, *Environmental pollution* **199**, 192 (2015).
- [66] I. O. Wong, C. Schooling, B. J. Cowling, and G. M. Leung, *British journal of cancer* **112**, 167 (2015).
- [67] P. Wu, A. M. Presanis, H. S. Bond, E. H. Lau, V. J. Fang, and B. J. Cowling, *Scientific reports* **7**, 929 (2017).
- [68] P. S. Dodds *et al.*, “Allotaxonomy for comparing complex systems with probability-based divergences,” (2020).