Allotaxonometry

Last updated: 2024/09/26, 08:35:21 EDT

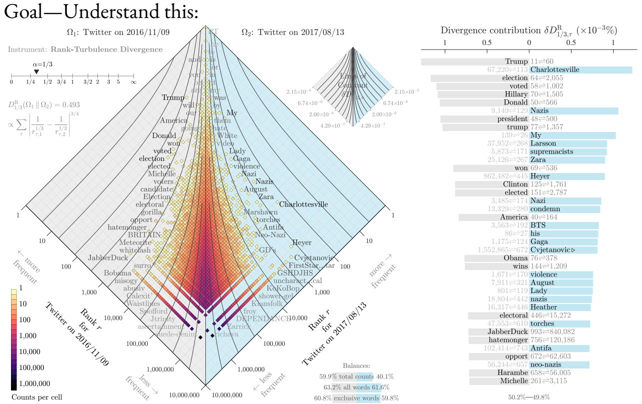
Principles of Complex Systems, Vols. 1, 2, & 3D
CSYS/MATH 6701, 6713, & a pretend number, 2024–2025

Prof. Peter Sheridan Dodds

Computational Story Lab | Vermont Complex Systems Center
Santa Fe Institute | University of Vermont

Outline

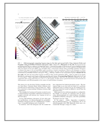
Goal—Understand this:



Site (papers, examples, code):
http://compstorylab.org/allotaxonometry/

Foundational papers:

"Allotaxonometry and rank-turbulence divergence: A universal instrument for comparing complex systems"
Dodds et al.,
EPJ Data Science, **12**, 1–42, 2023. [5]
EPJ Data Science version
arXiv version



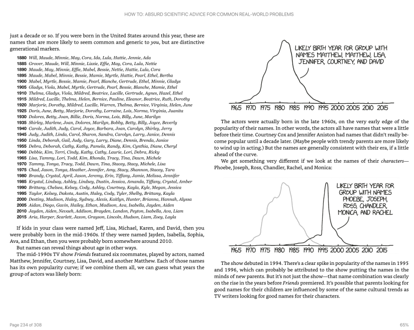
"Probability-turbulence divergence: A tunable allotaxonometric instrument for comparing heavy-tailed categorical distributions"
Dodds et al.,
, 2020. [6]

Basic science = Describe + Explain:

- Dashboards of single scale instruments helps us understand, monitor, and control systems.
- Archetype: Cockpit dashboard for flying a plane
- Okay if comprehendible.
- Complex systems present two problems for dashboards:
  1. Scale with internal diversity of components: We need meters for every species, every company, every word.
  2. Tracking change: We need to re-arrange meters on the fly.
- Goal—Create comprehendible, dynamically-adjusting, differential dashboards showing two pieces:[1]
  1. 'Big picture' map-like overview,
  2. A tunable ranking of components.

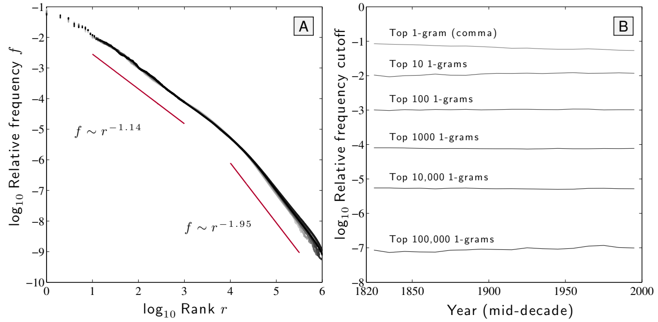
[1] See the lexicocalorimeter

Baby names, much studied: [12]



How to build a dynamical dashboard that helps sort through a massive number of interconnected time series?

"Is language evolution grinding to a halt? The scaling of lexical turbulence in English fiction suggests it is not"
Pechenick, Danforth, Dodds, Alshaabi, Adams, Reagan, Danforth, Frank, Reagan, and Danforth.
Journal of Computational Science, **21**, 24–37, 2017. [14]



For language, Zipf's law has two scaling regimes: [19]

$$f \sim \begin{cases} r^{-\alpha} \text{ for } r \ll r_b, \\ r^{-\alpha'} \text{ for } r \gg r_b, \end{cases}$$

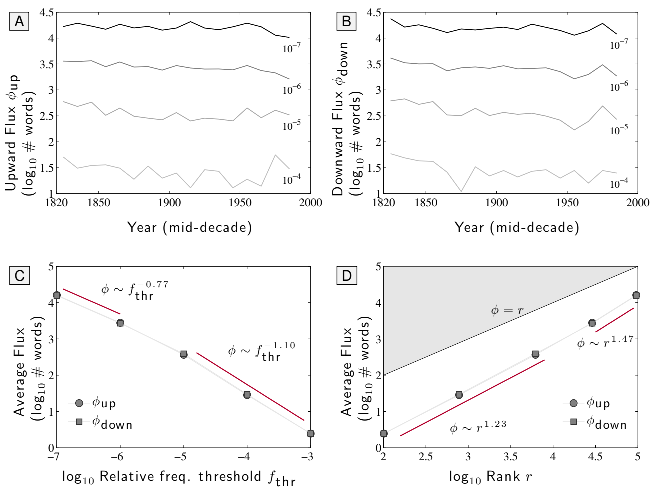
When comparing two texts, define Lexical turbulence as flux of words across a frequency threshold:

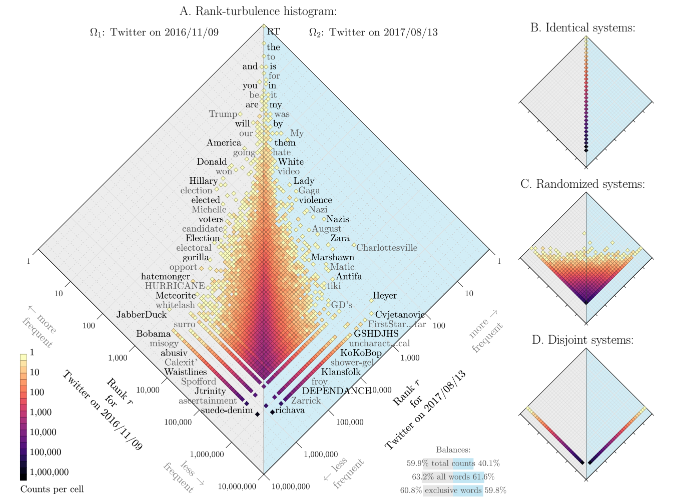
$$\phi \sim \begin{cases} f_{\text{thr}}^{-\mu} \text{ for } f_{\text{thr}} \ll f_b, \\ f_{\text{thr}}^{-\mu'} \text{ for } f_{\text{thr}} \gg f_b, \end{cases}$$

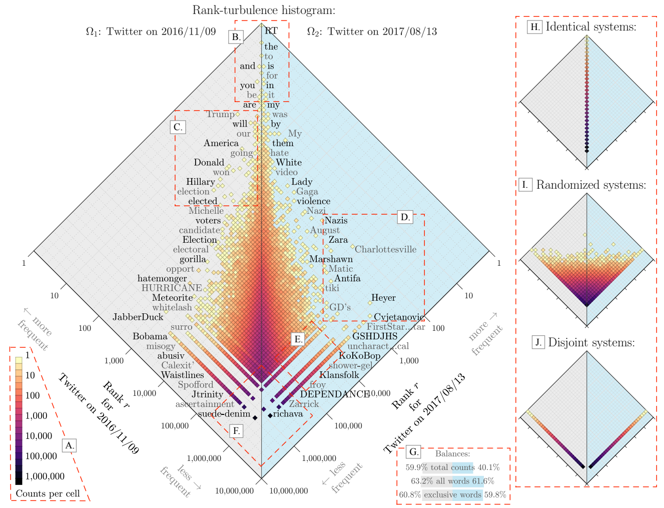
Estimates: $\mu \simeq 0.77$ and $\mu' \simeq 1.10$, and f_b is the scaling break point.

$$\phi \sim \begin{cases} r^{\nu} = r^{\alpha\mu'} \text{ for } r \ll r_b, \\ r^{\nu'} = r^{\alpha'\mu} \text{ for } r \gg r_b. \end{cases}$$

Estimates: Lower and upper exponents $\nu \simeq 1.23$ and $\nu' \simeq 1.47$.







Balances:

Top bar (optional)—Total size:

- Relative balance of system sizes.
- Examples: Total number of words in a book, total number of individuals in an ecology.

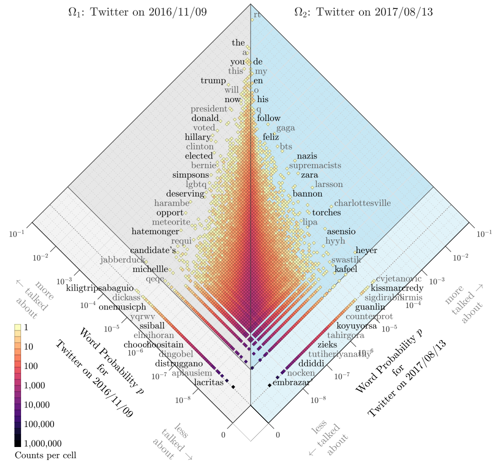
Middle bar—Types:

- Fraction of types in each system as a percentage of the union of types from both systems.

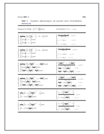
Bottom bar—Exclusive types:

- Types that are present in one system only are 'exclusive types'.
- $\Omega^{(1)}$-exclusive and $\Omega^{(2)}$-exclusive indicate which system an exclusive type belongs to.
- Percentage of exclusive types in a system relative to that system's total number of types.

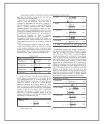
Probability-turbulence histogram:



So, so many ways to compare probability distributions:



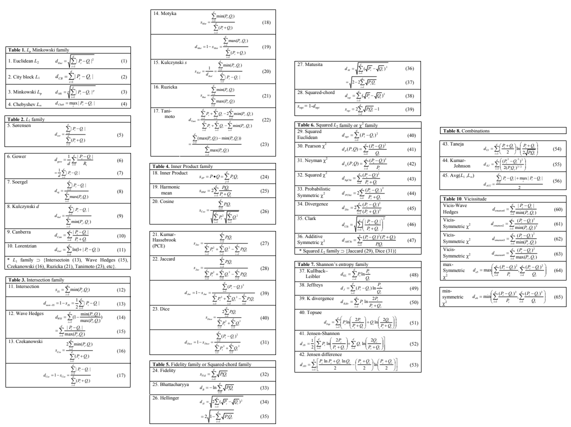
"Families of Alpha- Beta- and Gamma- Divergences: Flexible and Robust Measures of Similarities"
Cichocki and Amari,
Entropy, **12**, 1532-1568, 2010. [2]



"Comprehensive survey on distance/similarity measures between probability density functions"
Sung-Hyuk Cha,
International Journal of Mathematical Models and Methods in Applied Sciences, **1**, 300–307, 2007. [1]

- Comparisons are distances, divergences, similarities, inner products, fidelities ...
- 60ish kinds of comparisons grouped into 10 families
- A worry: Subsampled distributions with very heavy tails

Quite the festival:



Shannon tried to slow things down in 1956:

"The bandwagon"
Claude E Shannon,
IRE Transactions on Information Theory, **2**, 3, 1956. [16]

- "Information theory has ... become something of a scientific bandwagon."
- "While ... information theory is indeed a valuable tool ... [it] is certainly no panacea for the communication engineer or ... for anyone else.
- "A few first rate research papers are preferable to a large number that are poorly conceived or half-finished."

- We want two main things:
  1. A measure of difference between systems
  2. A way of sorting which types/species/words contribute to that difference
- For sorting, many comparisons give the same ordering.
- A few basic building blocks:
  - $|P_i - Q_i|$ (dominant)
  - $\max(P_i, Q_i)$
  - $\min(P_i, Q_i)$
  - $P_i Q_i$
  - $|P_i^{1/2} - Q_i^{1/2}|$ (Hellinger)

Table 1. L_p Minkowski family

1. Euclidean L_2	$d_{Euc} = \sqrt{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert^2}$	(1)
2. City block L_1	$d_{CB} = \sum_{i=1}^{d} \lvert P_i - Q_i \rvert$	(2)
3. Minkowski L_p	$d_{Mk} = \sqrt[p]{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert^p}$	(3)
4. Chebyshev L_∞	$d_{Cheb} = \max_i \lvert P_i - Q_i \rvert$	(4)

Table 2. L_1 family

5. Sørensen	$d_{sor} = \frac{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert}{\sum_{i=1}^{d} (P_i + Q_i)}$	(5)
6. Gower	$d_{gow} = \frac{1}{d}\sum_{i=1}^{d} \frac{\lvert P_i - Q_i \rvert}{R_i}$	(6)
	$= \frac{1}{d}\sum_{i=1}^{d} \lvert P_i - Q_i \rvert$	(7)
7. Soergel	$d_{sg} = \frac{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert}{\sum_{i=1}^{d} \max(P_i, Q_i)}$	(8)
8. Kulczynski d	$d_{kul} = \frac{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert}{\sum_{i=1}^{d} \min(P_i, Q_i)}$	(9)
9. Canberra	$d_{Can} = \sum_{i=1}^{d} \frac{\lvert P_i - Q_i \rvert}{P_i + Q_i}$	(10)
10. Lorentzian	$d_{Lor} = \sum_{i=1}^{d} \ln(1 + \lvert P_i - Q_i \rvert)$	(11)

* L_1 family ⊃ {Intersectoin (13), Wave Hedges (15), Czekanowski (16), Ruzicka (21), Tanimoto (23), etc}.

- Information theoretic sortings are more opaque
- No tunability

Table 1. L_p Minkowski family

1. Euclidean L_2	$d_{Euc} = \sqrt{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert^2}$	(1)
2. City block L_1	$d_{CB} = \sum_{i=1}^{d} \lvert P_i - Q_i \rvert$	(2)
3. Minkowski L_p	$d_{Mk} = \sqrt[p]{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert^p}$	(3)
4. Chebyshev L_∞	$d_{Cheb} = \max_i \lvert P_i - Q_i \rvert$	(4)

Table 2. L_1 family

5. Sørensen	$d_{sor} = \frac{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert}{\sum_{i=1}^{d} (P_i + Q_i)}$	(5)
6. Gower	$d_{gow} = \frac{1}{d}\sum_{i=1}^{d} \frac{\lvert P_i - Q_i \rvert}{R_i}$	(6)
	$= \frac{1}{d}\sum_{i=1}^{d} \lvert P_i - Q_i \rvert$	(7)
7. Soergel	$d_{sg} = \frac{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert}{\sum_{i=1}^{d} \max(P_i, Q_i)}$	(8)
8. Kulczynski d	$d_{kul} = \frac{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert}{\sum_{i=1}^{d} \min(P_i, Q_i)}$	(9)
9. Canberra	$d_{Can} = \sum_{i=1}^{d} \frac{\lvert P_i - Q_i \rvert}{P_i + Q_i}$	(10)
10. Lorentzian	$d_{Lor} = \sum_{i=1}^{d} \ln(1 + \lvert P_i - Q_i \rvert)$	(11)

* L_1 family ⊃ {Intersectoin (13), Wave Hedges (15), Czekanowski (16), Ruzicka (21), Tanimoto (23), etc}.

Shannon's Entropy:

$$H(P) = \langle \log_2 \frac{1}{p_\tau} \rangle = \sum_{\tau \in R_{1,2;\alpha}} p_\tau \log_2 \frac{1}{p_\tau} \quad (1)$$

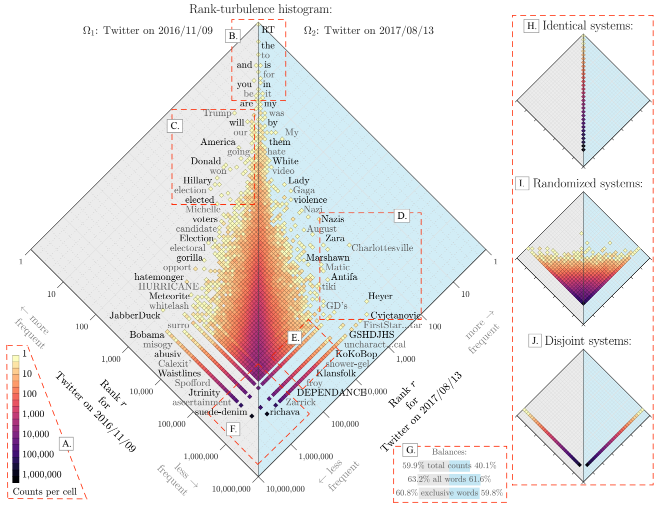
Kullback-Liebler (KL) divergence:

$$D^{\mathrm{KL}}(P_2 \,||\, P_1) = \left\langle \log_2 \frac{1}{p_{2,\tau}} - \log_2 \frac{1}{p_{1,\tau}} \right\rangle_{P_2}$$

$$= \sum_{\tau \in R_{1,2;\alpha}} p_{2,\tau} \left[\log_2 \frac{1}{p_{2,\tau}} - \log_2 \frac{1}{p_{1,\tau}} \right]$$

$$= \sum_{\tau \in R_{1,2;\alpha}} p_{2,\tau} \log_2 \frac{p_{1,\tau}}{p_{2,\tau}}. \quad (2)$$

- Problem: If just one component type in system 2 is not present in system 1, KL divergence = ∞.
- Solution: If we can't compare a spork and a platypus directly, we create a fictional spork-platypus hybrid.
- New problem: Re-read solution.



Some good things about ranks:

- Working with ranks is intuitive
- Affords some powerful statistics (e.g., Spearman's rank correlation coefficient)
- Can be used to generalize beyond systems with probabilities

A start:

$$\left| \frac{1}{r_{\tau,1}} - \frac{1}{r_{\tau,2}} \right|. \quad (5)$$

- Inverse of rank gives an increasing measure of 'importance'
- High rank means closer to rank 1
- We assign tied ranks for components of equal 'size'
- Issue: Biases toward high rank components

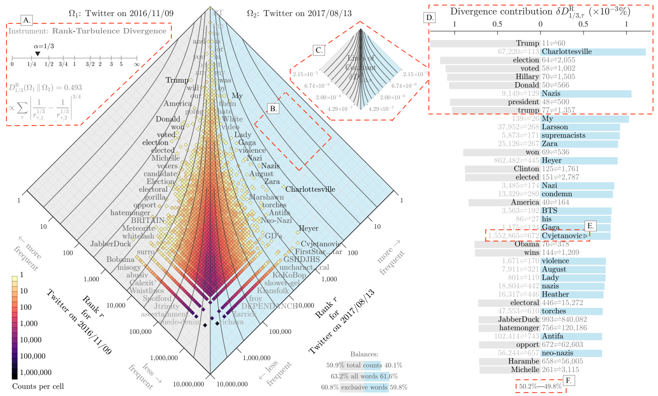
Jensen-Shannon divergence (JSD): [9, 7, 13, 1]

$$D^{\mathrm{JS}}(P_1 \,||\, P_2)$$
$$= \tfrac{1}{2} D^{\mathrm{KL}}\left(P_1 \,\middle\|\, \tfrac{1}{2}[P_1 + P_2]\right) + \tfrac{1}{2} D^{\mathrm{KL}}\left(P_2 \,\middle\|\, \tfrac{1}{2}[P_1 + P_2]\right)$$
$$= \frac{1}{2} \sum_{\tau \in R_{1,2;\alpha}} \left(p_{1,\tau} \log_2 \frac{p_{1,\tau}}{\frac{1}{2}[p_{1,\tau} + p_{2,\tau}]} + p_{2,\tau} \log_2 \frac{p_{2,\tau}}{\frac{1}{2}[p_{1,\tau} + p_{2,\tau}]} \right). \quad (3)$$

- Involving a third intermediate averaged system means JSD is now finite: $0 \le D^{\mathrm{JS}}(P_1 \,||\, P_2) \le 1$.
- Generalized entropy divergence: [2]

$$D^{\mathrm{AS2}}_\alpha(P_1 \,||\, P_2) =$$
$$\frac{1}{\alpha(\alpha-1)} \sum_{\tau \in R_{1,2;\alpha}} \left[\left(p_{\tau,1}^{1-\alpha} + p_{\tau,2}^{1-\alpha} \right) \left(\frac{p_{\tau,1} + p_{\tau,2}}{2} \right)^\alpha - (p_{\tau,1} + p_{\tau,2}) \right]. \quad (4)$$

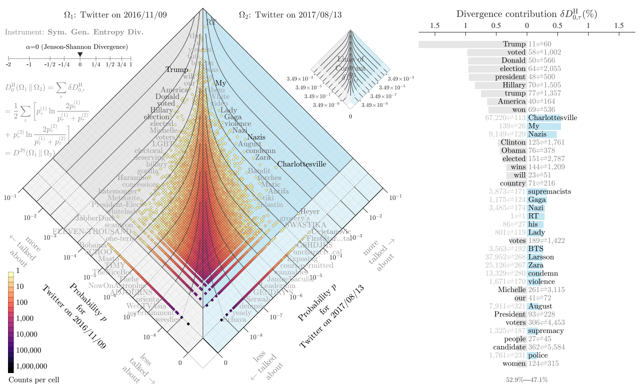
Produces JSD when $\alpha \to 0$.



We introduce a tuning parameter:

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

- As $\alpha \to 0$, high ranked components are increasingly dampened
- For words in texts, for example, the weight of common words and rare words move increasingly closer together.
- As $\alpha \to \infty$, high rank components will dominate.
- For texts, the contributions of rare words will vanish.



Desirable rank-turbulence divergence features:

1. Rank-based.
2. Symmetric.
3. Semi-positive: $D^{\mathrm{R}}_\alpha(\Omega_1 \,||\, \Omega_2) \ge 0$.
4. Linearly separable, for interpretability.
5. 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.).
6. Turbulence-handling: Suited for systems with rank-ordered component size distribution that are heavy-tailed.
7. Scalable: Allow for sensible comparisons across system sizes.
8. Tunable.
9. Story-finding: Features 1–8 combine to show which component types are most 'important'

Trouble:

- The limit of $\alpha \to 0$ does not behave well for

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

- The leading order term is:

$$\left(1 - \delta_{r_{\tau,1} r_{\tau,2}}\right) \alpha^{1/\alpha} \left| \ln \frac{r_{\tau,1}}{r_{\tau,2}} \right|^{1/\alpha}, \quad (7)$$

which heads toward ∞ as $\alpha \to 0$.

- Oops.
- But the insides look nutritious:

$$\left| \ln \frac{r_{\tau,1}}{r_{\tau,2}} \right|$$

is a nicely interpretable log-ratio of ranks.

Some reworking:

$$\delta D^{\mathrm{R}}_{\alpha,\tau}(R_1 \,\|\, R_2) \propto \frac{\alpha+1}{\alpha} \left| \frac{1}{\left[r_{\tau,1}\right]^{\alpha}} - \frac{1}{\left[r_{\tau,2}\right]^{\alpha}} \right|^{1/(\alpha+1)}. \quad (8)$$

- Keeps the core structure.
- Large α limit remains the same.
- $\alpha \to 0$ limit now returns log-ratio of ranks.
- Next: Sum over τ to get divergence.
- Still have an option for normalization.

Rank-turbulence divergence:

$$D^{\mathrm{R}}_{\alpha}(R_1 \,\|\, R_2) = \frac{1}{\mathcal{N}_{1,2;\alpha}} \sum_{\tau \in R_{1,2;\alpha}} \delta D^{\mathrm{R}}_{\alpha,\tau}(R_1 \,\|\, R_2) \quad (9)$$

Normalization:

- Take a data-driven rather than analytic approach to determining $\mathcal{N}_{1,2;\alpha}$.
- Compute $\mathcal{N}_{1,2;\alpha}$ by taking the two systems to be disjoint while maintaining their underlying Zipf distributions.
- Ensures: $0 \le D^{\mathrm{R}}_{\alpha}(R_1 \,\|\, R_2) \le 1$
- Limits of 0 and 1 correspond to the two systems having identical and disjoint Zipf distributions.

Rank-turbulence divergence:

Summing over all types, dividing by a normalization prefactor $\mathcal{N}_{1,2;\alpha}$ we have our prototype:

$$D^{\mathrm{R}}_{\alpha}(R_1 \,\|\, R_2) = \frac{1}{\mathcal{N}_{1,2;\alpha}} \frac{\alpha+1}{\alpha} \sum_{\tau \in R_{1,2;\alpha}} \left| \frac{1}{\left[r_{\tau,1}\right]^{\alpha}} - \frac{1}{\left[r_{\tau,2}\right]^{\alpha}} \right|^{1/(\alpha+1)}. \quad (10)$$

General normalization:

- Iif 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:

$$\mathcal{N}_{1,2;\alpha} = \frac{\alpha+1}{\alpha} \sum_{\tau \in R_1} \left| \frac{1}{\left[r_{\tau,1}\right]^{\alpha}} - \frac{1}{\left[N_1 + \frac{1}{2}N_2\right]^{\alpha}} \right|^{1/(\alpha+1)} + \frac{\alpha+1}{\alpha} \sum_{\tau \in R_2} \left| \frac{1}{\left[N_2 + \frac{1}{2}N_1\right]^{\alpha}} - \frac{1}{\left[r_{\tau,2}\right]^{\alpha}} \right|^{1/(\alpha+1)}. \quad (11)$$

Limit of $\alpha \to 0$:

$$D^{\mathrm{R}}_{0}(R_1 \,\|\, R_2) = \sum_{\tau \in R_{1,2;\alpha}} \delta D^{\mathrm{R}}_{0,\tau} = \frac{1}{\mathcal{N}_{1,2;0}} \sum_{\tau \in R_{1,2;\alpha}} \left| \ln \frac{r_{\tau,1}}{r_{\tau,2}} \right|, \quad (12)$$

where

$$\mathcal{N}_{1,2;0} = \sum_{\tau \in R_1} \left| \ln \frac{r_{\tau,1}}{N_1 + \frac{1}{2}N_2} \right| + \sum_{\tau \in R_2} \left| \ln \frac{r_{\tau,2}}{\frac{1}{2}N_1 + N_2} \right|. \quad (13)$$

- Largest rank ratios dominate.

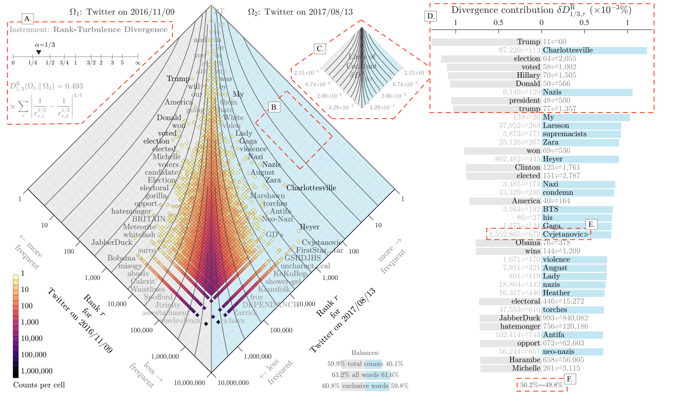
Limit of $\alpha \to \infty$:

$$D^{\mathrm{R}}_{\infty}(R_1 \,\|\, R_2) = \sum_{\tau \in R_{1,2;\alpha}} \delta D^{\mathrm{R}}_{\infty,\tau} = \frac{1}{\mathcal{N}_{1,2;\infty}} \sum_{\tau \in R_{1,2;\alpha}} \left(1 - \delta_{r_{\tau,1} r_{\tau,2}}\right) \max_{\tau} \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\}. \quad (14)$$

where

$$\mathcal{N}_{1,2;\infty} = \sum_{\tau \in R_1} \frac{1}{r_{\tau,1}} + \sum_{\tau \in R_2} \frac{1}{r_{\tau,2}}. \quad (15)$$

- Highest ranks dominate.



Probability-turbulence divergence:

$$D^{\mathrm{P}}_{\alpha}(P_1 \,\|\, P_2) = \frac{1}{\mathcal{N}^{\mathrm{P}}_{1,2;\alpha}} \frac{\alpha+1}{\alpha} \sum_{\tau \in R_{1,2;\alpha}} \left| \left[p_{\tau,1}\right]^{\alpha} - \left[p_{\tau,2}\right]^{\alpha} \right|^{1/(\alpha+1)}. \quad (16)$$

- For the unnormalized version ($\mathcal{N}^{\mathrm{P}}_{1,2;\alpha}$=1), some troubles return with 0 probabilities and $\alpha \to 0$.
- Weep not: $\mathcal{N}^{\mathrm{P}}_{1,2;\alpha}$ will save the day.

Normalization:

With no matching types, the probability of a type present in one system is zero in the other, and the sum can be split between the two systems' types:

$$\mathcal{N}^{\mathrm{P}}_{1,2;\alpha} = \frac{\alpha+1}{\alpha} \sum_{\tau \in R_1} \left[p_{\tau,1}\right]^{\alpha/(\alpha+1)} + \frac{\alpha+1}{\alpha} \sum_{\tau \in R_2} \left[p_{\tau,2}\right]^{\alpha/(\alpha+1)} \quad (17)$$

Limit of $\alpha = 0$ for probability-turbulence divergence

- if both $p_{\tau,1} > 0$ and $p_{\tau,2} > 0$ then

$$\lim_{\alpha \to 0} \frac{\alpha+1}{\alpha} \left| \left[p_{\tau,1}\right]^{\alpha} - \left[p_{\tau,2}\right]^{\alpha} \right|^{1/(\alpha+1)} = \left| \ln \frac{p_{\tau,2}}{p_{\tau,1}} \right|. \quad (18)$$

- But if $p_{\tau,1} = 0$ or $p_{\tau,2} = 0$, limit diverges as $1/\alpha$.

Limit of α=0 for probability-turbulence divergence

- Normalization:

$$\mathcal{N}^{\mathrm{P}}_{1,2;\alpha} \rightarrow \frac{1}{\alpha}\left(N_1 + N_2\right). \quad (19)$$

- Because the normalization also diverges as $1/\alpha$, the divergence will be zero when there are no exclusive types and non-zero when there are exclusive types.

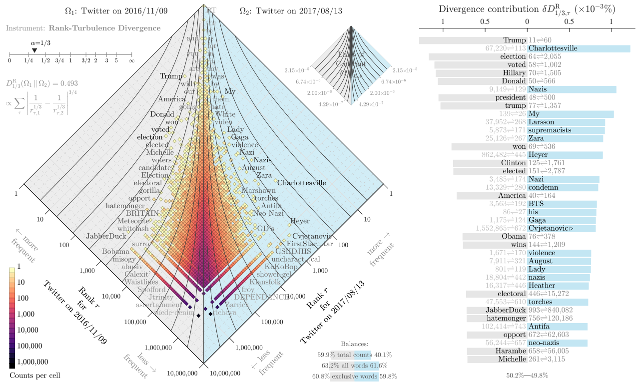
Combine these cases into a single expression:

$$D^{\mathrm{P}}_{0}(P_1 \,\|\, P_2) = \frac{1}{(N_1 + N_2)} \sum_{\tau \in R_{1,2;0}} \left(\delta_{p_{\tau,1},0} + \delta_{0,p_{\tau,2}} \right). \quad (20)$$

- The term $\left(\delta_{p_{\tau,1},0} + \delta_{0,p_{\tau,2}}\right)$ returns 1 if either $p_{\tau,1} = 0$ or $p_{\tau,2} = 0$, and 0 otherwise when both $p_{\tau,1} > 0$ and $p_{\tau,2} > 0$.
- Ratio of types that are exclusive to one system relative to the total possible such types,

Type contribution ordering for the limit of α=0

- In terms of contribution to the divergence score, all exclusive types supply a weight of $1/(N_1 + N_2)$. We can order them by preserving their ordering as $\alpha \rightarrow 0$, which amounts to ordering by descending probability in the system in which they appear.
- And while types that appear in both systems make no contribution to $D^{\mathrm{P}}_{0}(P_1 \,\|\, P_2)$, we can still order them according to the log ratio of their probabilities.
- The overall ordering of types by divergence contribution for α=0 is then: (1) exclusive types by descending probability and then (2) types appearing in both systems by descending log ratio.

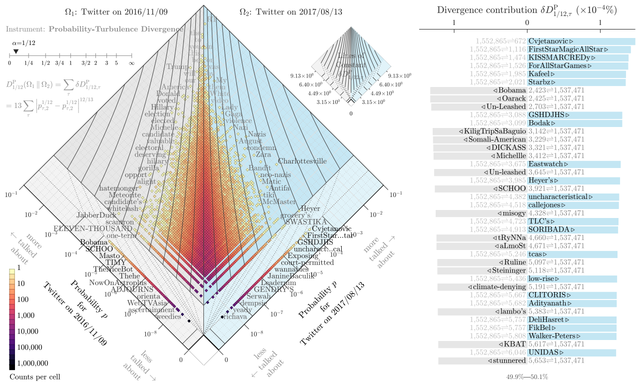


Limit of α=∞ for probability-turbulence divergence

$$D^{\mathrm{P}}_{\infty}(P_1 \,\|\, P_2) = \frac{1}{2} \sum_{\tau \in R_{1,2;\infty}} \left(1 - \delta_{p_{\tau,1},p_{\tau,2}}\right) \max\left(p_{\tau,1}, p_{\tau,2}\right) \quad (21)$$

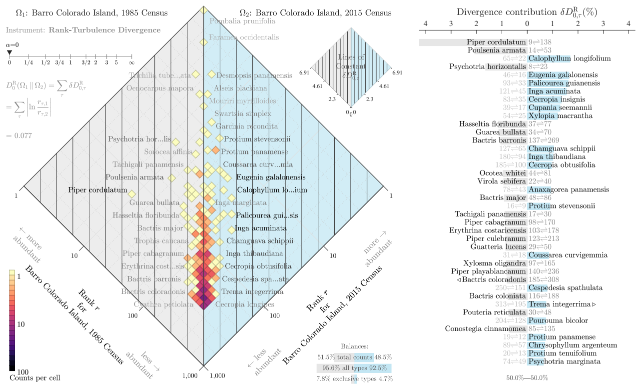
where

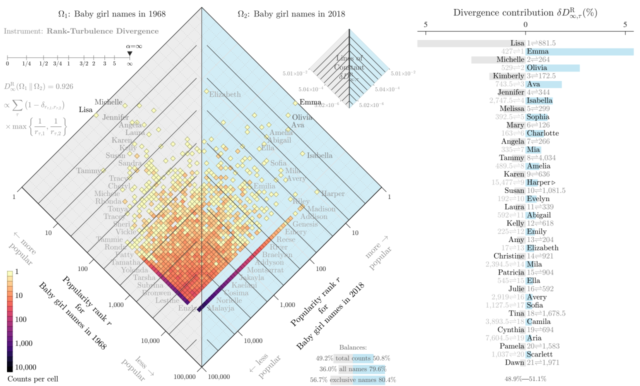
$$\mathcal{N}^{\mathrm{P}}_{1,2;\infty} = \sum_{\tau \in R_{1,2;\infty}} \left(p_{\tau,1} + p_{\tau,2} \right) = 1 + 1 = 2. \quad (22)$$

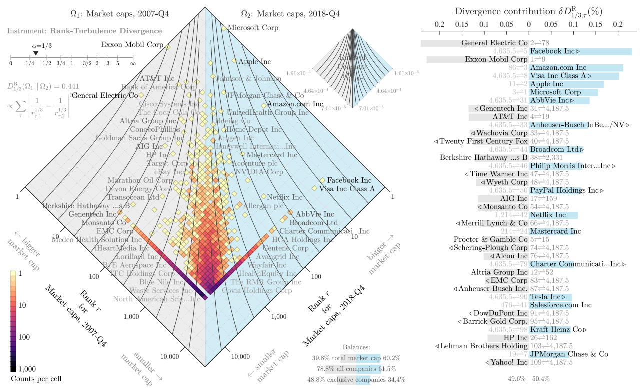


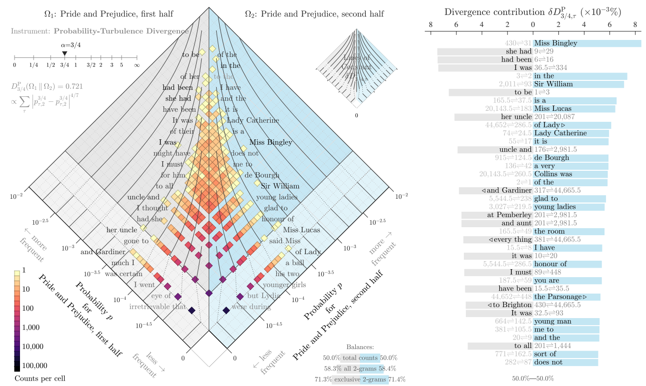
Connections for PTD:

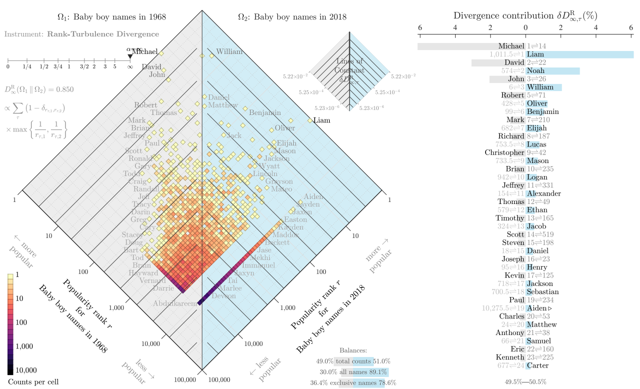
- $\alpha = 0$: Similarity measure Sørensen-Dice coefficient [4, 17, 10], F_1 score of a test's accuracy [18, 15].
- $\alpha = 1/2$: Hellinger distance [8] and Mautusita distance [11].
- $\alpha = 1$: Many including all $L^{(p)}$-norm type constructions.
- $\alpha = \infty$: Motyka distance [3].

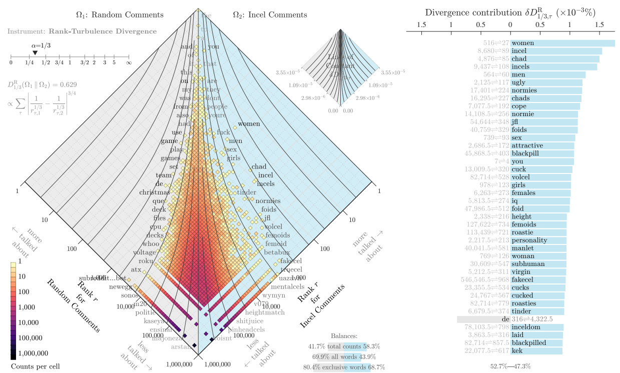












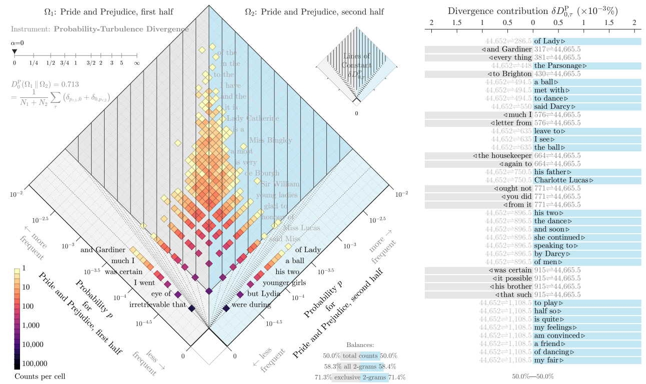
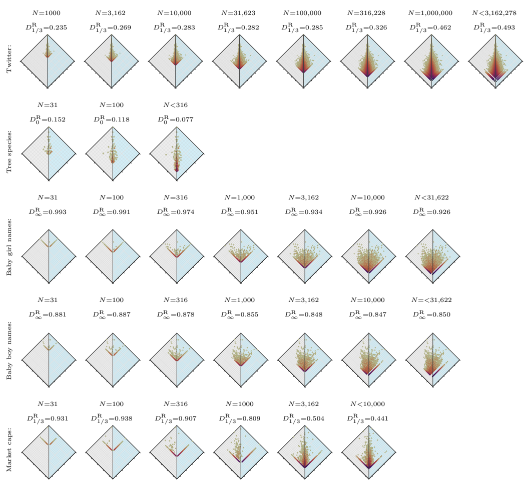
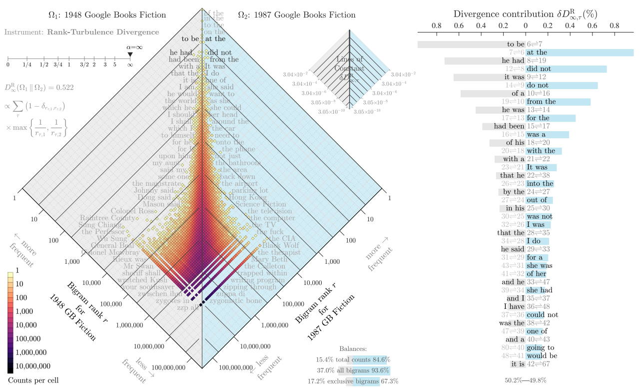
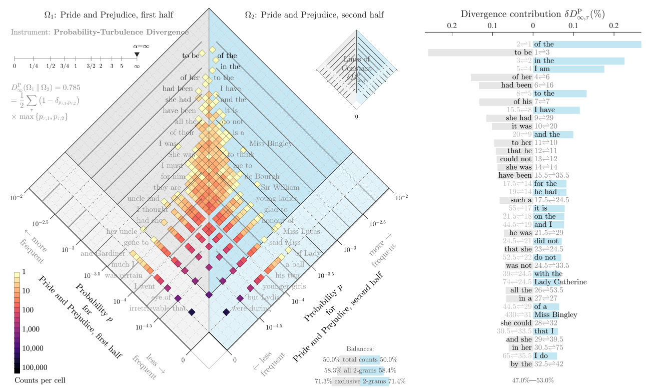


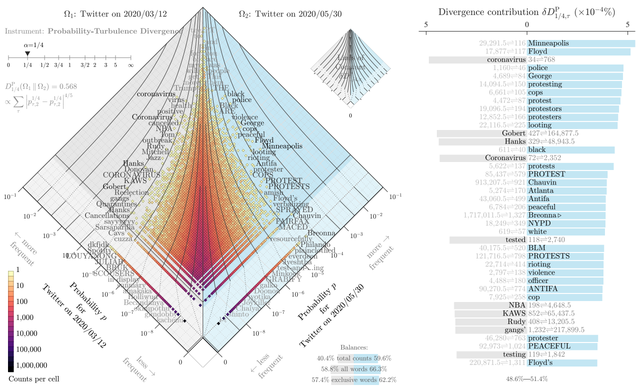
FIG. 8. **Rank-turbulence divergence allotaxonograph [34] of word rank distributions in the incel vs random comment corpora.** The rank-rank histogram on the left shows the density of words by their rank in the incel comments corpus against their rank in the random comments corpus. Words at the top of the diamond are higher frequency, or lower rank. For example, the word "the" appears at the highest observed frequency, and thus has the lowest rank, 1. This word has the lowest rank in both corpora, so its coordinates lie along the center vertical line in the plot. Words such as "women" diverge from the center line because their rank in the incel corpus is higher than in the random corpus. The top 40 words with greatest divergence contribution are shown on the right. In this comparison, nearly all of the top 40 words are more common in the incel corpus, so they point to the right. The word that has the most notable change in rank from the random to incel corpus is "women", the object of hatred

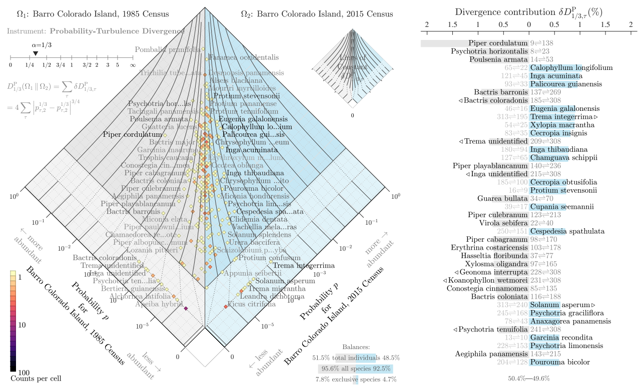
Effect of subsampling:





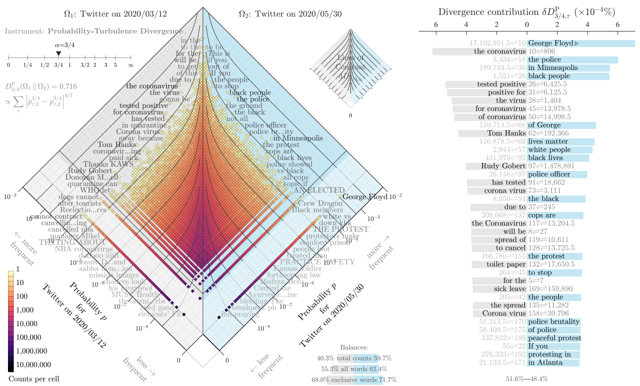






Flipbooks for PTD:

- Jane Austen:
  - Pride and Prejudice, 1-grams
  - Pride and Prejudice, 2-grams
  - Pride and Prejudice, 3-grams
- Social media:
  - Twitter, 1-grams
  - Twitter, 2-grams
  - Twitter, 3-grams
- Ecology:
  - Barro Colorado Island

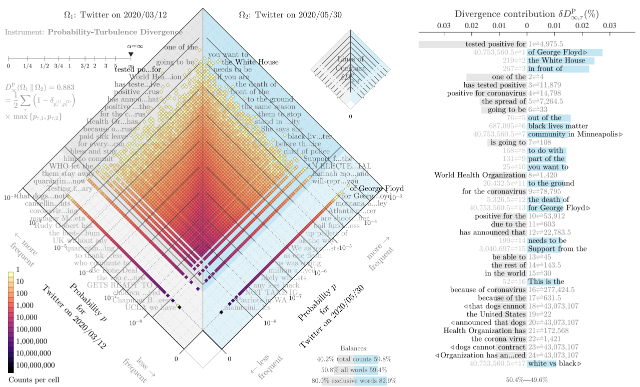


Flipbooks for RTD:

- Twitter:
  - allotaxonometer-flipbook-1-rank-div.pdf
  - allotaxonometer-flipbook-2-probability-div.pdf
  - allotaxonometer-flipbook-3-gen-entropy-div.pdf
- Market caps:
  - allotaxonometer-flipbook-4-marketcaps-6years-rank-div.pdf
- Baby names:
  - allotaxonometer-flipbook-5-babynames-girls-50years-rank-div.pdf
  - allotaxonometer-flipbook-6-babynames-boys-50years-rank-div.pdf
  - Baby girl names over time relative to 1950
  - Baby boy names over time relative to 1950

Code:

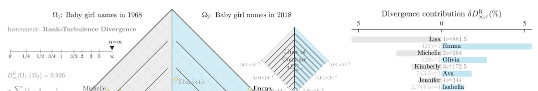
https://gitlab.com/compstorylab/allotaxonometer



- Google books:
  - allotaxonometer-flipbook-7-google-books-onegrams-rank-div.pdf
  - allotaxonometer-flipbook-8-google-books-bigrams-rank-div.pdf
  - allotaxonometer-flipbook-9-google-books-trigrams-rank-div.pdf

Claims, exaggerations, reminders:

- Needed for comparing large-scale complex systems: Comprehendible, dynamically-adjusting, differential dashboards.
- Many measures seem poorly motivated and largely unexamined (e.g., JSD).
- Of value: Combining big-picture maps with ranked lists.
- Online tunable versions of rank-turbulence divergence now exist:
  - App version: https://allotaxp.vercel.app/
  - Observable version: https://observablehq.com/@jstonge/allotaxonometer-4-all
  - Github: https://github.com/jstonge/allotaxp
- Future: Probability-turbulence divergence plus many other instruments.



References I

[1] S.-H. Cha.
Comprehensive survey on distance/similarity measures between probability density functions.
International Journal of Mathematical Models and Methods in Applied Sciences, 1:300–307, 2007. pdf

[2] A. Cichocki and S.-i. Amari.
Families of Alpha- Beta- and Gamma- divergences: Flexible and robust measures of similarities.
Entropy, 12:1532–1568, 2010. pdf

[3] M.-M. Deza and E. Deza.
Dictionary of Distances.
Elsevier, 2006.

References II

[4] L. R. Dice.
Measures of the amount of ecologic association between species.
Ecology, 26:297–302, 1945.

[5] P. S. Dodds, J. R. Minot, M. V. Arnold, T. Alshaabi, J. L. Adams, D. R. Dewhurst, T. J. Gray, M. R. Frank, A. J. Reagan, and C. M. Danforth.
Allotaxonometry and rank-turbulence divergence: A universal instrument for comparing complex systems.
EPJ Data Science, 12(37):1–42, 2023.
Available online at https://arxiv.org/abs/2002.09770. pdf

References III

[6] P. S. Dodds, J. R. Minot, M. V. Arnold, T. Alshaabi, J. L. Adams, A. J. Reagan, and C. M. Danforth.
Probability-turbulence divergence: A tunable allotaxonometric instrument for comparing heavy-tailed categorical distributions, 2020.
Available online at https://arxiv.org/abs/2008.13078. pdf

[7] D. M. Endres and J. E. Schindelin.
A new metric for probability distributions.
IEEE Transactions on Information theory, 2003. pdf

[8] E. Hellinger.
Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen.
Journal für die reine und angewandte Mathematik (Crelles Journal), 1909(136):210–271, 1909. pdf

References IV

[9] J. Lin.
Divergence measures based on the Shannon entropy.
IEEE Transactions on Information theory, 37(1):145–151, 1991. pdf

[10] J. Looman and J. B. Campbell.
Adaptation of Sørensen's k (1948) for estimating unit affinities in prairie vegetation.
Ecology, 41(3):409–416, 1960. pdf

[11] K. Matusita et al.
Decision rules, based on the distance, for problems of fit, two samples, and estimation.
The Annals of Mathematical Statistics, 26(4):631–640, 1955. pdf

References V

[12] R. Munroe.
How To: Absurd Scientific Advice for Common Real-World Problems.
Penguin, 2019.

[13] F. Osterreicher and I. Vajda.
A new class of metric divergences on probability spaces and its applicability in statistics.
Annals of the Institute of Statistical Mathematics, 55(3):639–653, 2003.

[14] E. A. Pechenick, C. M. Danforth, and P. S. Dodds.
Is language evolution grinding to a halt? The scaling of lexical turbulence in English fiction suggests it is not.
Journal of Computational Science, 21:24–37, 2017. pdf

[15] Y. Sasaki.
The truth of the *f*-measure, 2007.

References VI

[16] C. E. Shannon.
The bandwagon.
IRE Transactions on Information Theory, 2(1):3, 1956. pdf

[17] T. Sorensen.
A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons.
Videnski Selskab Biologiske Skrifter, 5:1–34, 1948.

[18] C. J. Van Rijsbergen.
Information retrieval.
Butterworth-Heinemann, 2nd edition, 1979.

References VII

[19] J. R. Williams, J. P. Bagrow, C. M. Danforth, and P. S. Dodds.
Text mixing shapes the anatomy of rank-frequency distributions.
Physical Review E, 91:052811, 2015. pdf