

Allotaxonomy


Last updated: 2024/09/26, 08:35:43 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



Licensed under the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) 

The PoCVerse
Allotaxonomy
1 of 73

A plenitude of
distances

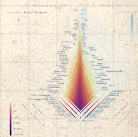
Rank-turbulence
divergence

Probability-
turbulence divergence

Explorations

Nutshell

References



These slides are brought to you by:

Sealie & Lambie
Productions



The PoCverse
Allotaxonomy
2 of 73

A plenitude of
distances

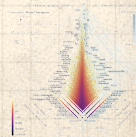
Rank-turbulence
divergence

Probability-
turbulence divergence

Explorations

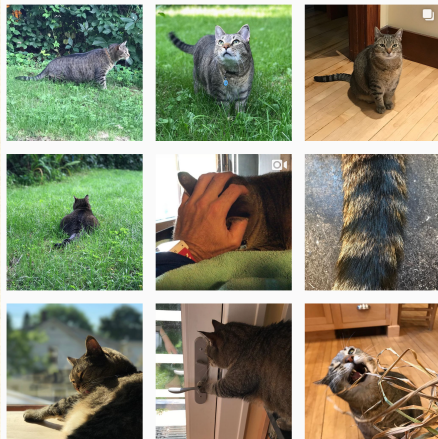
Nutshell



References



These slides are also brought to you by:

Special Guest Executive Producer



 On Instagram at [pratchett_the_cat](https://www.instagram.com/pratchett_the_cat) 

The PoCverse
Allotaxonomy
3 of 73

A plenitude of
distances

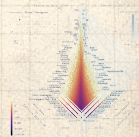
Rank-turbulence
divergence

Probability-
turbulence divergence

Explorations

Nutshell

References



Outline

A plenitude of distances

Rank-turbulence divergence

Probability-turbulence divergence

Explorations

Nutshell

References

The PoCSverse
Allotaxonomy
4 of 73

A plenitude of
distances

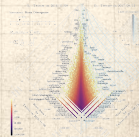
Rank-turbulence
divergence

Probability-
turbulence divergence

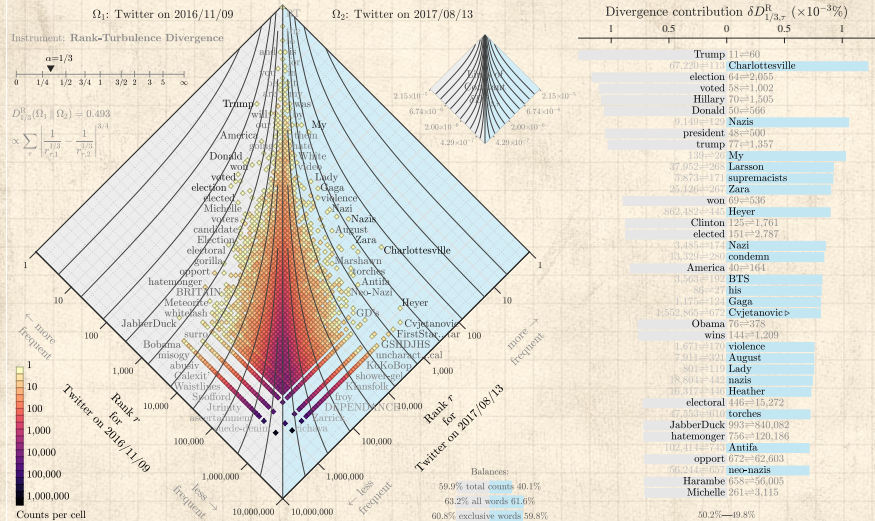
Explorations

Nutshell

References



Goal—Understand this:



A plenitude of
distances

Rank-turbulence
divergence

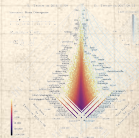
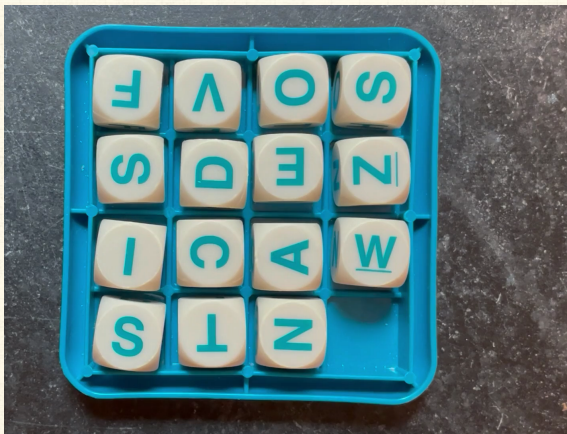
Probability-
turbulence divergence

Explorations

Nutshell

References

The Boggoracle Speaks:




Site (papers, examples, code):

<http://compstorylab.org/allotaxonomy/> 


Foundational papers:



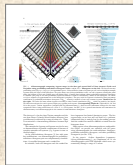
“Allotaxonomy and rank-turbulence divergence:
A universal instrument for comparing complex
systems” 


Dodds et al.,

EPJ Data Science, **12**, 1–42, 2023. ^[5]

[EPJ Data Science version](#) 

[arXiv version](#) 





“Probability-turbulence divergence: A tunable
allotaxonomic 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

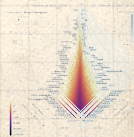
 Okay if comprehensible.


 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 comprehensible, dynamically-adjusting, differential dashboards showing two pieces:¹

1. ‘Big picture’ map-like overview,
2. A tunable ranking of components.



¹See the [lexicocalorimeter](#) 

Baby names, much studied: ^[12]

just a decade or so. If you were born in the United States around this year, these are names that are more likely to seem common and generic to you, but are distinctive generational markers.

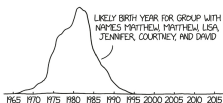
1800 Will, Minnie, Minnie, May, Cora, Ida, Lulu, Hattie, Annie, Ada
1805 Greer, Maude, Will, Minnie, Lottie, Effie, May, Cora, Lulu, Nellie
1890 Maude, May, Minnie, Effie, Maebel, Bessie, Nettie, Hattie, Lulu, Cora
1905 Maude, Maebel, Minnie, Bessie, Minnie, Hattie, Hattie, Pearl, Ethel, Bertha
1900 Maebel, Myrtle, Bessie, Minnie, Pearl, Blanche, Gertrude, Ethel, Minnie, Gladys
1905 Gladys, Viola, Maebel, Myrtle, Gertrude, Pearl, Bessie, Blanche, Mammie, Ethel
1910 Thelma, Gladys, Viola, Mildred, Beatrice, Lucille, Gertrude, Agnes, Hazel, Ethel
1915 Mildred, Lucille, Thelma, Helen, Bernice, Pauline, Eleanor, Beatrice, Ruth, Dorothy
1920 Marjorie, Dorothy, Mildred, Lucille, Warren, Thelma, Bernice, Virginia, Helen, June
1925 Doris, Jane, Betty, Marjorie, Dorothy, Lorraine, Lulu, Norma, Virginia, Juvenile
1930 Dolores, Betty, Joan, Billie, Doris, Norma, Lois, Billy, Joan, Marilyn
1935 Shirley, Marjorie, Joan, Dolores, Marilyn, Bobby, Betty, Billy, Joyce, Beverly
1940 Corale, Judith, Judy, Carol, Joyce, Barbara, Joan, Carolyn, Shirley, Jerry
1945 Judy, Judith, Linda, Carol, Sharon, Sandra, Carolyn, Larry, Janice, Dennis
1950 Linda, Deborah, Gail, Andy, Gary, Larry, Diane, Beverly, Brenda, Annie
1955 Debra, Deborah, Cathy, Kathy, Pamela, Randy, Kim, Cynthia, Diane, Cheryl
1960 Debbie, Kim, Terri, Cindy, Kathy, Cathy, Laurie, Lori, Debra, Ricky
1965 Lisa, Tammy, Lori, Todd, Kim, Rhonda, Tracy, Tina, Anne, Michele
1970 Tammy, Tanya, Tracy, Todd, Dawn, Tina, Sherry, Stacy, Michele, Lisa
1975 Chad, Susan, Tanya, Heather, Jennifer, Amy, Sherry, Shannon, Sherry, Tara
1980 Amanda, Crystal, April, Jason, Jeremy, Erin, Tiffany, Jamie, Melissa, Jennifer
1985 Crystal, Lindsay, Ashley, Lindsay, Doreen, Austin, Amanda, Tiffany, Crystal, Amber
1990 Brittany, Chelsea, Kelsey, Cindy, Ashley, Courtney, Kayla, Kyle, Megan, Jessica
1995 Taylor, Kelsey, Dakota, Austin, Haley, Cindy, Tyler, Shelby, Brittany, Kayla
2000 Destiny, Madison, Emily, Sydney, Haley, Hunter, Britney, Hannah, Alyssa
2005 Aidan, Diego, Gavin, Hailey, Ethan, Madison, Ava, Isabella, Aiyden, Aiden
2010 Jayden, Aiden, Noah, Addison, Brayden, Peyton, Isabella, Ava, Liam
2015 Ari, Harper, Scarlett, Sloan, Graysen, Livielle, Hudson, Liam, Zoey, Layla

If kids in your class were named Jeff, Lisa, Michael, Karen, and David, then you were probably born in the mid-1960s. If they were named Jayden, Isabella, Sophia, Ava, and Ethan, then you were probably born somewhere around 2010.

But names can reveal things about age in other ways.

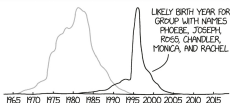
The mid-1990s TV show *Friends* featured six roommates, played by actors, named Matthew, Jennifer, Courtney, Lisa, David, and another Matthew. Each of those names has its own popularity curve; if we combine them all, we can guess what years the group of actors was likely born.

HOW TO: ABSURD SCIENTIFIC ADVICE FOR COMMON REAL-WORLD PROBLEMS



The actors were actually born in the late 1960s, on the very early edge of the popularity of their names. In other words, the actors all have names that were a little before their time. Courtney Cox and Jennifer Aniston had names that didn't really become popular until a decade later. (Maybe people with trendy parents are more likely to wind up in acting.) But the names are generally consistent with their era, if a little ahead of the curve.

We get something very different if we look at the names of their characters—Phoebe, Joseph, Ross, Chandler, Rachel, and Monica:



The show debuted in 1994. There's a clear spike in popularity of the names in 1995 and 1996, which can probably be attributed to the show putting the names in the minds of new parents. But it's not just the show—that name combination was clearly on the rise in the years before *Friends* premiered. It's possible that parents looking for good names for their children are influenced by some of the same cultural trends as TV writers looking for good names for their characters.

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

The PoCVerse
Allotaxonomy
9 of 73

A plenitude of
distances

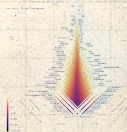
Rank-turbulence
divergence

Probability-
turbulence divergence

Explorations

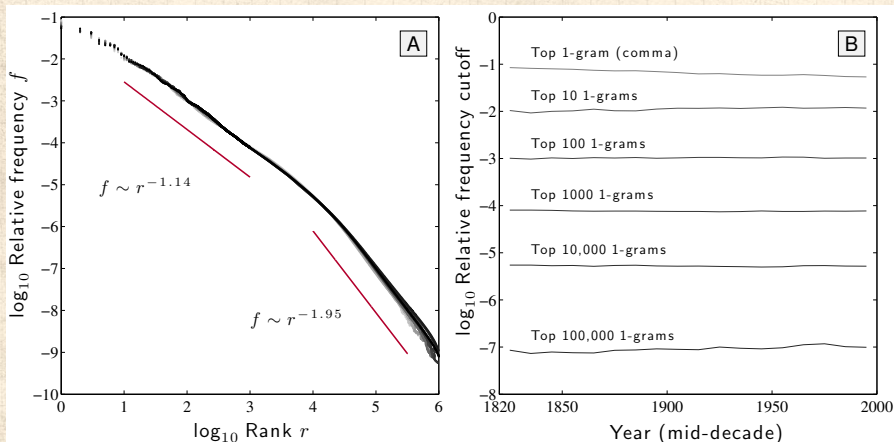
Nutshell

References





“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]

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

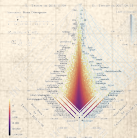
When comparing two texts, define Lexical turbulence as flux of words across a frequency threshold:

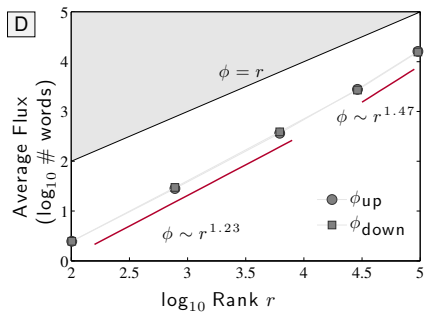
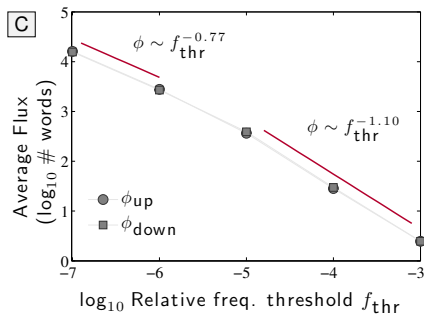
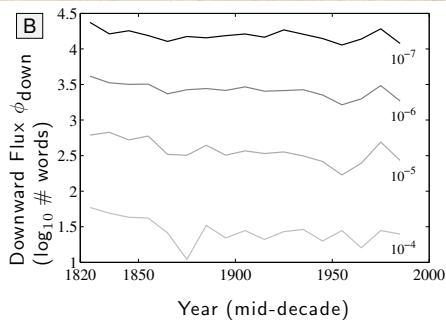
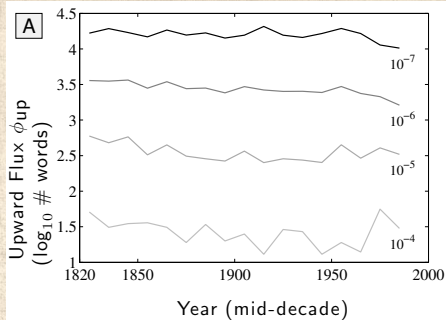
$$\phi \sim \begin{cases} f_{\text{thr}}^{-\mu} & \text{for } f_{\text{thr}} \ll f_b, \\ f_{\text{thr}}^{-\mu'} & \text{for } f_{\text{thr}} \gg f_b, \end{cases}$$

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 $\nu' \simeq 1.47$.

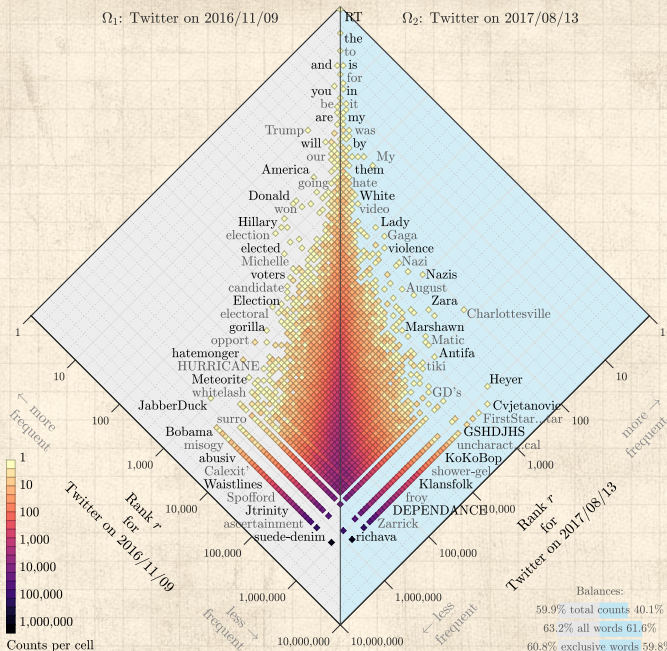




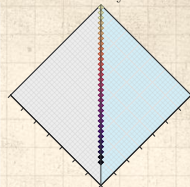
A. Rank-turbulence histogram:

Ω_1 : Twitter on 2016/11/09

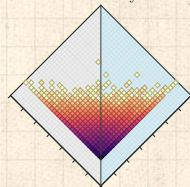
Ω_2 : Twitter on 2017/08/13



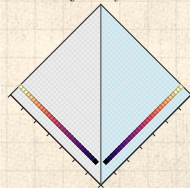
B. Identical systems:



C. Randomized systems:



D. Disjoint systems:



Balances:

59.9% total counts 40.1%

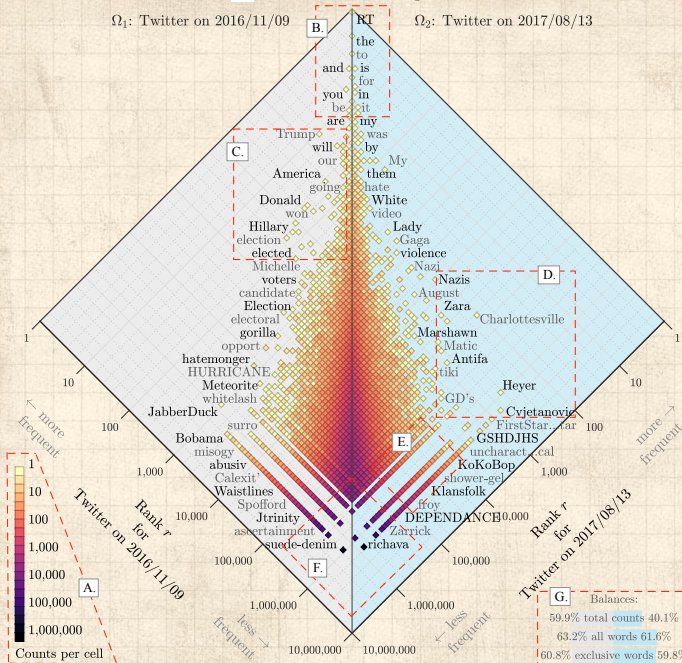
63.2% all words 61.6%

60.8% exclusive words 59.8%

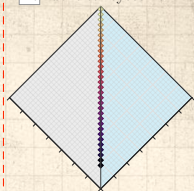
Rank-turbulence histogram:

Ω_1 : Twitter on 2016/11/09

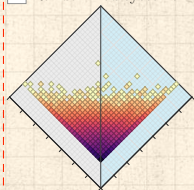
Ω_2 : Twitter on 2017/08/13



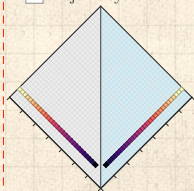
H. Identical systems:



I. Randomized systems:





J. Disjoint systems:




G. Balances:
 59.9% total counts 40.1%
 63.2% all words 61.6%
 60.8% exclusive words 59.8%

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:

The PoCServe
Allotaxonomy
16 of 73

A plenitude of
distances

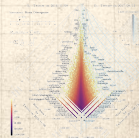
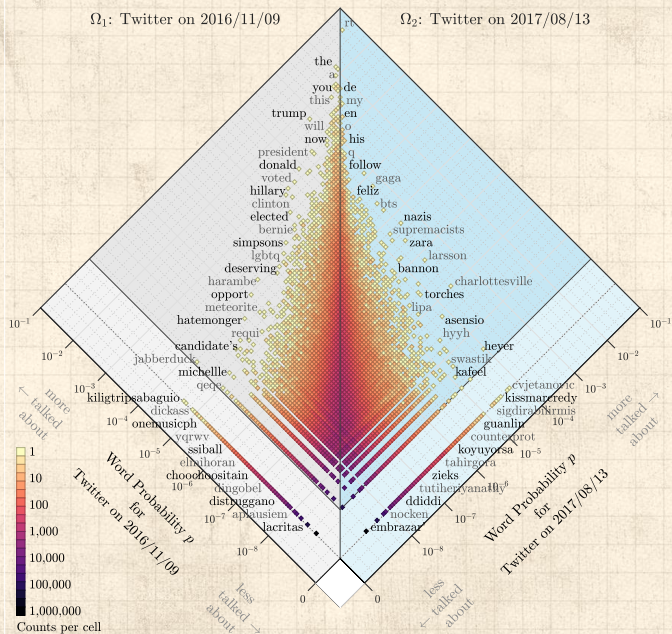
Rank-turbulence
divergence

Probability-
turbulence divergence

Explorations

Nutshell

References



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]



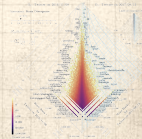
Comparisons are distances, divergences [↗](#), similarities, inner products, fidelities ...



60ish kinds of comparisons grouped into 10 families



A worry: Subsampled distributions with very heavy tails



Quite the festival:

Table 1. L_r Minkowski family

1. Euclidean L_2	$d_{L_2} = \sqrt{\frac{1}{2}(\ P-Q\ ^2)}$ (1)
2. City block L_1	$d_{L_1} = \frac{1}{2}(\ P-Q\)$ (2)
3. Minkowski L_r	$d_{L_r} = \sqrt[r]{\frac{1}{2}(\ P-Q\ ^r)}$ (3)
4. Chebyshev L_∞	$d_{L_\infty} = \max\{\ P-Q\ \}$ (4)

Table 2. L_r family

5. Somenzi	$d_{L_r} = \frac{\sum \ P-Q\ }{\sum \ P+Q\ }$ (5)
------------	---

6. Gower	$d_{L_r} = \frac{1}{2} \sum \sqrt[r]{\frac{\ P-Q\ }{\ P+Q\ }}$ (6)
----------	--

	$+ \frac{1}{2} \sum \sqrt[r]{\ P-Q\ }$ (7)
--	--

7. Szeged	$d_{L_r} = \frac{\sum \ P-Q\ }{\sum \max\{P, Q\}}$ (8)
-----------	--

8. Kulczyński d'	$d_{L_r} = \frac{\sum \ P-Q\ }{\sum \max\{P, Q\}}$ (9)
--------------------	--

9. Canberra	$d_{L_r} = \frac{\sum \sqrt[r]{\ P-Q\ }}{\sum \ P+Q\ }$ (10)
-------------	--

10. Lorentzian	$d_{L_r} = \sum \ln(1+\ P-Q\)$ (11)
----------------	--------------------------------------

* L_r family \supset Intersection (13), Wave Hedges (15), Czekanowski (16), Ruzicka (21), Tamimoto (23), etc.

Table 3. Intersection family

11. Intersection	$x_{\cap} = \sum \min\{P, Q\}$ (12)
------------------	-------------------------------------

	$d_{\cap} = 1 - x_{\cap} = \frac{1}{2} \sum \ P-Q\ $ (13)
--	---

12. Wave Hedges	$d_{\cap} = \sum \left(\frac{\min\{P, Q\}}{\max\{P, Q\}} \right)$ (14)
-----------------	---

	$= \sum \sqrt{\frac{\ P-Q\ }{\max\{P, Q\}}}$ (15)
--	---

13. Czekanowski	$x_{\cap} = \frac{\sum \min\{P, Q\}}{\sum \ P+Q\ }$ (16)
-----------------	--

	$d_{\cap} = 1 - x_{\cap} = \frac{\sum \ P-Q\ }{\sum \ P+Q\ }$ (17)
--	--

	$d_{\cap} = 1 - x_{\cap} = \frac{\sum \ P-Q\ }{\sum \ P+Q\ }$ (17)
--	--

	$d_{\cap} = 1 - x_{\cap} = \frac{\sum \ P-Q\ }{\sum \ P+Q\ }$ (17)
--	--

	$d_{\cap} = 1 - x_{\cap} = \frac{\sum \ P-Q\ }{\sum \ P+Q\ }$ (17)
--	--

	$d_{\cap} = 1 - x_{\cap} = \frac{\sum \ P-Q\ }{\sum \ P+Q\ }$ (17)
--	--

	$d_{\cap} = 1 - x_{\cap} = \frac{\sum \ P-Q\ }{\sum \ P+Q\ }$ (17)
--	--

	$d_{\cap} = 1 - x_{\cap} = \frac{\sum \ P-Q\ }{\sum \ P+Q\ }$ (17)
--	--

	$d_{\cap} = 1 - x_{\cap} = \frac{\sum \ P-Q\ }{\sum \ P+Q\ }$ (17)
--	--

	$d_{\cap} = 1 - x_{\cap} = \frac{\sum \ P-Q\ }{\sum \ P+Q\ }$ (17)
--	--

14. Motyka	$d_{L_r} = \frac{\sum \max\{P, Q\}}{\sum \ P+Q\ }$ (18)
------------	---

	$d_{L_r} = 1 - x_{\cap} = \frac{\sum \max\{P, Q\}}{\sum \ P+Q\ }$ (19)
--	--

15. Kulczyński r	$x_{L_r} = \frac{1}{d_{L_r}} \frac{\sum \min\{P, Q\}}{\sum \ P-Q\ }$ (20)
--------------------	---

16. Ruzicka	$x_{L_r} = \frac{\sum \max\{P, Q\}}{\sum \max\{P, Q\}}$ (21)
-------------	--

17. Tamimoto	$d_{L_r} = \frac{\sum \ P-Q\ - \sum \min\{P, Q\}}{\sum \ P+Q\ - \sum \max\{P, Q\}}$ (22)
--------------	--

	$= \frac{\sum (\max\{P, Q\} - \min\{P, Q\})}{\sum \max\{P, Q\}}$ (23)
--	---

Table 4. Inner Product family

18. Inner Product	$x_{\cap} = P+Q - \sum P_i Q_i$ (24)
-------------------	--------------------------------------

19. Harmonic mean	$x_{\cap} = \sum \frac{P_i Q_i}{P_i + Q_i}$ (25)
-------------------	--

20. Cosine	$x_{\cap} = \frac{\sum P_i Q_i}{\sqrt{\sum P_i^2} \sqrt{\sum Q_i^2}}$ (26)
------------	--

21. Kumar-Hauschok (PCE)	$x_{\cap} = \frac{\sum P_i Q_i}{\sum P_i^r + \sum Q_i^r - \sum P_i Q_i}$ (27)
--------------------------	---

22. Jaccard	$x_{\cap} = \frac{\sum P_i Q_i}{\sum P_i^r + \sum Q_i^r - \sum P_i Q_i}$ (28)
-------------	---

	$d_{\cap} = 1 - x_{\cap} = \frac{\sum \ P-Q\ ^r}{\sum P_i^r + \sum Q_i^r - \sum P_i Q_i}$ (29)
--	--

23. Dice	$x_{\cap} = \frac{\sum P_i Q_i}{\sum P_i^2 + \sum Q_i^2}$ (30)
----------	--

	$d_{\cap} = 1 - x_{\cap} = \frac{\sum \ P-Q\ ^2}{\sum P_i^2 + \sum Q_i^2}$ (31)
--	---

Table 5. Fidelity family or Squared-chord family

24. Fidelity	$x_{\cap} = \sum \sqrt{P_i Q_i}$ (32)
--------------	---------------------------------------

25. Bhattacharyya	$d_{\cap} = \ln \sum \sqrt{P_i Q_i}$ (33)
-------------------	---

26. Hellinger	$d_{\cap} = \sqrt{\frac{1}{2} \sum \sqrt{P_i^2 + Q_i^2}}$ (34)
---------------	--

	$= \sqrt{\frac{1}{2} \sum \sqrt{P_i Q_i}}$ (35)
--	---

27. Matusita	$d_{\cap} = \sqrt{\frac{1}{2} \sum \sqrt{P_i^2 + Q_i^2}}$ (36)
--------------	--

	$= \sqrt{\frac{1}{2} \sum \sqrt{P_i Q_i}}$ (37)
--	---

28. Squared-chord	$d_{\cap} = \sum \sqrt{P_i^2 + Q_i^2}$ (38)
-------------------	---

	$x_{\cap} = 1 - d_{\cap} = 2 \sum \sqrt{P_i Q_i} - 1$ (39)
--	--

Table 6. Squared L_r family or χ^2 family

29. Squared Euclidean	$d_{\cap} = \sum (P_i - Q_i)^2$ (40)
-----------------------	--------------------------------------

30. Pearson χ^2	$d_{\cap}(P, Q) = \sum \frac{(P_i - Q_i)^2}{Q_i}$ (41)
----------------------	--

31. Neyman χ^2	$d_{\cap}(P, Q) = \sum \frac{(P_i - Q_i)^2}{P_i}$ (42)
---------------------	--

32. Squared χ^2	$d_{\cap} = \sum \frac{(P_i - Q_i)^2}{P_i + Q_i}$ (43)
----------------------	--

33. Probabilistic Symmetric χ^2	$d_{\cap} = \sum \frac{(P_i - Q_i)^2}{P_i + Q_i}$ (44)
--------------------------------------	--

34. Divergence	$d_{\cap} = 2 \sum \frac{(P_i - Q_i)^2}{(P_i + Q_i)}$ (45)
----------------	--

35. Clark	$d_{\cap} = \sqrt{\sum \left(\frac{P_i - Q_i}{P_i + Q_i} \right)^2}$ (46)
-----------	--

36. Additive Symmetric χ^2	$d_{\cap} = \sum \frac{(P_i - Q_i)^2 (P_i + Q_i)}{P_i Q_i}$ (47)
---------------------------------	--

* Squared L_r family \supset Jaccard (28), Dice (31)

Table 7. Shannon's entropy family

37. Kullback-Libler	$d_{\cap} = \sum P_i \ln \frac{P_i}{Q_i}$ (48)
---------------------	--

38. Jeffreys	$d_{\cap} = \sum (P_i - Q_i) \ln \frac{P_i}{Q_i}$ (49)
--------------	--

39. K. divergence	$d_{\cap} = \sum P_i \ln \frac{2P_i}{P_i + Q_i}$ (50)
-------------------	---

40. Topoc	$d_{\cap} = \sum P_i \ln \left(\frac{2P_i}{P_i + Q_i} \right) + Q_i \ln \left(\frac{2Q_i}{P_i + Q_i} \right)$ (51)
-----------	--

41. Jensen-Shannon	$d_{\cap} = \frac{1}{2} \sum P_i \ln \left(\frac{2P_i}{P_i + Q_i} \right) + \frac{1}{2} \sum Q_i \ln \left(\frac{2Q_i}{P_i + Q_i} \right)$ (52)
--------------------	---

42. Jensen divergence	$d_{\cap} = \frac{1}{2} \sum \left[P_i \ln \frac{P_i + Q_i}{2} \ln \left(\frac{P_i + Q_i}{2} \right) + Q_i \ln \left(\frac{P_i + Q_i}{2} \right) \right]$ (53)
-----------------------	---

Table 8. Combinations

43. Taneja	$d_{\cap} = \frac{\sum (\ P+Q\)}{2} \ln \left(\frac{\ P-Q\ }{\sqrt{P_i Q_i}} \right)$ (54)
------------	--

44. Kumar-Johnson	$d_{\cap} = \sum \left(\frac{\ P-Q\ ^r}{2(P_i Q_i)^r} \right)$ (55)
-------------------	--

45. Avg(L, L _r)	$d_{\cap} = \frac{\sum \ P-Q\ + \max\{P-Q\}}{2}$ (56)
-----------------------------	--

Table 10. Vicinostale

Vicin-Wave Hedges	$d_{\cap} = \sum \frac{\ P-Q\ }{\max\{P, Q\}}$ (60)
-------------------	---

Vicin-Symmetric χ^2	$d_{\cap} = \sum \frac{(P_i - Q_i)^2}{\max\{P, Q\}}$ (61)
--------------------------	---

Vicin-Symmetric χ^2	$d_{\cap} = \sum \frac{(P_i - Q_i)^2}{\max\{P, Q\}}$ (62)
--------------------------	---

Vicin-Symmetric χ^2	$d_{\cap} = \sum \frac{(P_i - Q_i)^2}{\max\{P, Q\}}$ (63)
--------------------------	---

max-symmetric	$d_{\cap} = \max \left(\sum \frac{(P_i - Q_i)^2}{P_i}, \sum \frac{(P_i - Q_i)^2}{Q_i} \right)$ (64)
---------------	--

min-symmetric	$d_{\cap} = \min \left(\sum \frac{(P_i - Q_i)^2}{P_i}, \sum \frac{(P_i - Q_i)^2}{Q_i} \right)$ (65)
---------------	--

A plentitude of distances

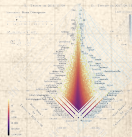
Rank-turbulence divergence

Probability-turbulence divergence

Explorations


Nutshell

References



Shannon tried to slow things down in 1956:



“The bandwagon” 

Claude E Shannon,

IRE Transactions on Information Theory, 2, 3,
1956. ^[16]



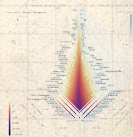
“Information theory has ... become something of a scientific bandwagon.”



“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.”



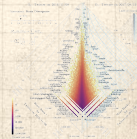


Table 1. L_p Minkowski family		
1. Euclidean L_2	$d_{Euc} = \sqrt{\sum_{i=1}^d P_i - Q_i ^2}$	(1)
2. City block L_1	$d_{CB} = \sum_{i=1}^d P_i - Q_i $	(2)
3. Minkowski L_p	$d_{Mk} = \sqrt[p]{\sum_{i=1}^d P_i - Q_i ^p}$	(3)
4. Chebyshev L_∞	$d_{Cheb} = \max_i P_i - Q_i $	(4)

Table 2. L_1 family		
5. Sørensen	$d_{sor} = \frac{\sum_{i=1}^d P_i - Q_i }{\sum_{i=1}^d (P_i + Q_i)}$	(5)

6. Gower	$d_{gow} = \frac{1}{d} \sum_{i=1}^d \frac{ P_i - Q_i }{R_i}$	(6)
	$= \frac{1}{d} \sum_{i=1}^d P_i - Q_i $	(7)

7. Soergel	$d_{sg} = \frac{\sum_{i=1}^d P_i - Q_i }{\sum_{i=1}^d \max(P_i, Q_i)}$	(8)
------------	---	-----

8. Kulczynski d	$d_{kul} = \frac{\sum_{i=1}^d P_i - Q_i }{\sum_{i=1}^d \min(P_i, Q_i)}$	(9)
-------------------	--	-----

9. Canberra	$d_{can} = \sum_{i=1}^d \frac{ P_i - Q_i }{P_i + Q_i}$	(10)
-------------	--	------

10. Lorentzian	$d_{lor} = \sum_{i=1}^d \ln(1 + P_i - Q_i)$	(11)
----------------	---	------

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



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

Table 2. L_1 family

5. Sørensen	$d_{sor} = \frac{\sum_{i=1}^d P_i - Q_i }{\sum_{i=1}^d (P_i + Q_i)}$	(5)
-------------	---	-----

6. Gower	$d_{gow} = \frac{1}{d} \sum_{i=1}^d \frac{ P_i - Q_i }{R_i}$	(6)
----------	--	-----

$$= \frac{1}{d} \sum_{i=1}^d |P_i - Q_i| \quad (7)$$

7. Soergel	$d_{sg} = \frac{\sum_{i=1}^d P_i - Q_i }{\sum_{i=1}^d \max(P_i, Q_i)}$	(8)
------------	---	-----

8. Kulczynski d	$d_{kul} = \frac{\sum_{i=1}^d P_i - Q_i }{\sum_{i=1}^d \min(P_i, Q_i)}$	(9)
-------------------	--	-----

9. Canberra	$d_{can} = \sum_{i=1}^d \frac{ P_i - Q_i }{P_i + Q_i}$	(10)
-------------	--	------

10. Lorentzian	$d_{Lor} = \sum_{i=1}^d \ln(1 + P_i - Q_i)$	(11)
----------------	---	------

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

The PoCverse
Allotaxonomy
21 of 73

A plenitude of
distances

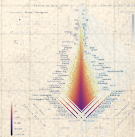
Rank-turbulence
divergence

Probability-
turbulence divergence

Explorations

Nutshell

References



Information theoretic
sortings are more opaque



No tunability

Shannon's Entropy:

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

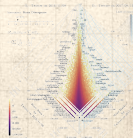
Kullback-Liebler (KL) divergence:

$$\begin{aligned} D^{\text{KL}}(P_2 \parallel P_1) &= \left\langle \log_2 \frac{1}{p_{2,\tau}} - \log_2 \frac{1}{p_{1,\tau}} \right\rangle_{P_2} \\ &= \sum_{\tau \in R_{1,2;\alpha}} p_{2,\tau} \left[\log_2 \frac{1}{p_{2,\tau}} - \log_2 \frac{1}{p_{1,\tau}} \right] \\ &= \sum_{\tau \in R_{1,2;\alpha}} p_{2,\tau} \log_2 \frac{p_{1,\tau}}{p_{2,\tau}}. \end{aligned} \quad (2)$$

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.



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

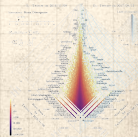
$$\begin{aligned}
 D^{\text{JS}}(P_1 \parallel P_2) &= \frac{1}{2} D^{\text{KL}}\left(P_1 \parallel \frac{1}{2}[P_1 + P_2]\right) + \frac{1}{2} D^{\text{KL}}\left(P_2 \parallel \frac{1}{2}[P_1 + P_2]\right) \\
 &= \frac{1}{2} \sum_{\tau \in R_{1,2;\alpha}} \left(p_{1,\tau} \log_2 \frac{p_{1,\tau}}{\frac{1}{2}[p_{1,\tau} + p_{2,\tau}]} + p_{2,\tau} \log_2 \frac{p_{2,\tau}}{\frac{1}{2}[p_{1,\tau} + p_{2,\tau}]} \right). \tag{3}
 \end{aligned}$$

🗑️ Involving a third intermediate averaged system means JSD is now finite:
 $0 \leq D^{\text{JS}}(P_1 \parallel P_2) \leq 1.$

🗑️ Generalized entropy divergence: [2]

$$\begin{aligned}
 D_{\alpha}^{\text{AS2}}(P_1 \parallel P_2) &= \\
 \frac{1}{\alpha(\alpha-1)} \sum_{\tau \in R_{1,2;\alpha}} &\left[(p_{\tau,1}^{1-\alpha} + p_{\tau,2}^{1-\alpha}) \left(\frac{p_{\tau,1} + p_{\tau,2}}{2} \right)^{\alpha} - (p_{\tau,1} + p_{\tau,2}) \right]. \tag{4}
 \end{aligned}$$

Produces JSD when $\alpha \rightarrow 0.$



Ω_1 : Twitter on 2016/11/09

Ω_2 : Twitter on 2017/08/13

Divergence contribution $\delta D_{0,r}^H$ (%)

Instrument: Sym. Gen. Entropy Div.

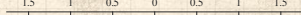
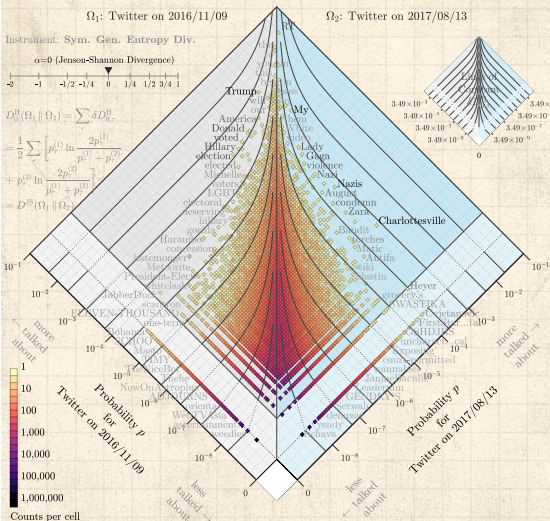
$\alpha=0$ (Jenson-Shannon Divergence)

$$D_{0,r}^H(\Omega_1 || \Omega_2) = \sum \delta D_{0,r}^H$$

$$= \frac{1}{2} \sum_r \left[p_r^{(1)} \ln \frac{2p_r^{(1)}}{p_r^{(1)} + p_r^{(2)}} \right]$$

$$+ p_r^{(2)} \ln \frac{2p_r^{(2)}}{p_r^{(1)} + p_r^{(2)}}$$

$$= D^{JS}(\Omega_1 || \Omega_2)$$



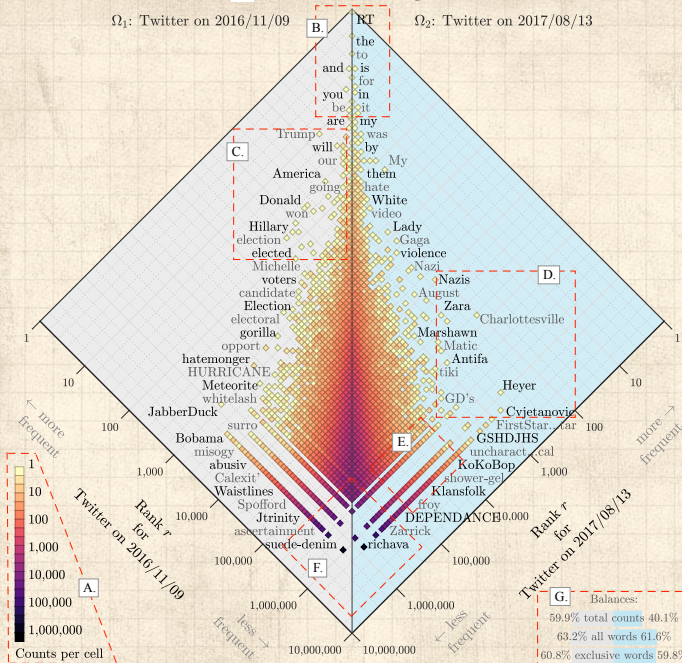
Trump	11=60
voted	58=1,002
Donald	50=566
election	64=2,055
president	48=500
Hillary	70=1,505
trump	77=1,357
America	40=164
won	69=536
67,220=113	Charlotteville
139=20	My
9,149=129	Nazis
Clinton	125=1,761
Obama	76=378
elected	151=2,787
wins	144=1,209
will	23=51
country	71=216
5,873=171	supremacists
1,175=124	Gaga
3,485=174	Nazi
1=1	RT
86=27	his
801=119	Lady
votes	180=1,422
3,563=192	BTS
37,952=268	Larsson
25,126=267	Zara
13,329=280	condemn
1,671=170	violence
Michelle	261=3,115
our	41=72
7,911=321	August
President	93=228
voters	306=4,453
1,325=187	supremacy
people	27=45
candidate	362=5,584
1,761=231	police
women	124=315

52.9%—47.1%

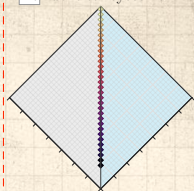
Rank-turbulence histogram:

Ω_1 : Twitter on 2016/11/09

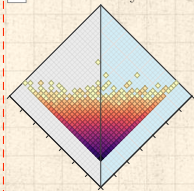
Ω_2 : Twitter on 2017/08/13



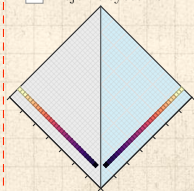
H. Identical systems:



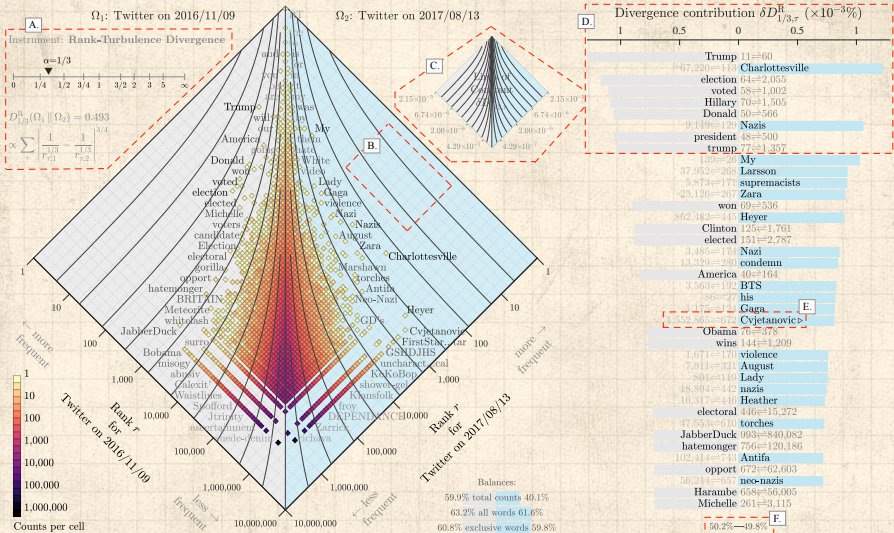
I. Randomized systems:



J. Disjoint systems:

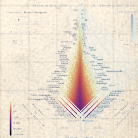


G. Balances:
 59.9% total counts 40.1%
 63.2% all words 61.6%
 60.8% exclusive words 59.8%



Desirable rank-turbulence divergence features:

1. Rank-based.
2. Symmetric.
3. Semi-positive: $D_{\alpha}^R(\Omega_1 \parallel \Omega_2) \geq 0$.
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’



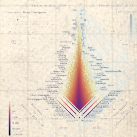
Some good things about ranks:

- 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:

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

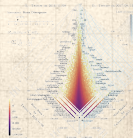
- Inverse of rank gives an increasing measure of 'importance'
- High rank means closer to rank 1
- We assign tied ranks for components of equal 'size'
- Issue: Biases toward high rank components



We introduce a tuning parameter:

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

- As $\alpha \rightarrow 0$, 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 $\alpha \rightarrow \infty$, high rank components will dominate.
- For texts, the contributions of rare words will vanish.



Trouble:

☰ The limit of $\alpha \rightarrow 0$ 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:

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

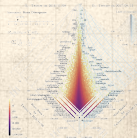
which heads toward ∞ as $\alpha \rightarrow 0$.

☰ Oops.

☰ But the insides look nutritious:






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

is a nicely interpretable log-ratio of ranks.



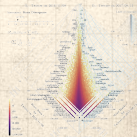
Some reworking:

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

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

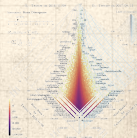
Rank-turbulence divergence:

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



Normalization:

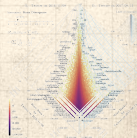
- Take a data-driven rather than analytic approach to determining $\mathcal{N}_{1,2;\alpha}$.
- Compute $\mathcal{N}_{1,2;\alpha}$ by taking the two systems to be disjoint while maintaining their underlying Zipf distributions.
- Ensures: $0 \leq D_{\alpha}^R(R_1 \parallel R_2) \leq 1$
- Limits of 0 and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

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



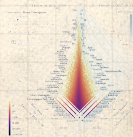
General normalization:

✚ 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:

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




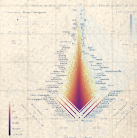
Limit of $\alpha \rightarrow 0$:

$$D_0^R(R_1 \parallel R_2) = \sum_{\tau \in R_{1,2;\alpha}} \delta D_{0,\tau}^R = \frac{1}{\mathcal{N}_{1,2;0}} \sum_{\tau \in R_{1,2;\alpha}} \left| \ln \frac{r_{\tau,1}}{r_{\tau,2}} \right|, \quad (12)$$

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 $\alpha \rightarrow \infty$:

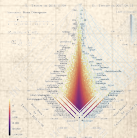
$$\begin{aligned} D_{\infty}^R(R_1 \| R_2) &= \sum_{\tau \in R_{1,2;\alpha}} \delta D_{\infty,\tau}^R \\ &= \frac{1}{\mathcal{N}_{1,2;\infty}} \sum_{\tau \in R_{1,2;\alpha}} (1 - \delta_{r_{\tau,1} r_{\tau,2}}) \max_{\tau} \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\}. \end{aligned} \quad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

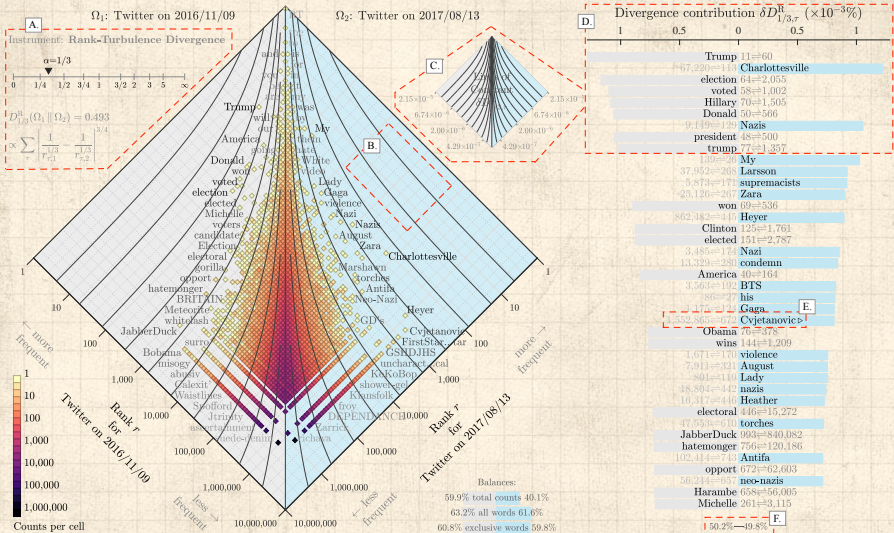
where

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




Highest ranks dominate.






Probability-turbulence divergence:

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

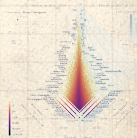
 For the unnormalized version ($\mathcal{N}_{1,2;\alpha}^{\text{P}}=1$), some troubles return with 0 probabilities and $\alpha \rightarrow 0$.

 Weep not: $\mathcal{N}_{1,2;\alpha}^{\text{P}}$ 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:

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

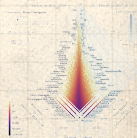


Limit of $\alpha = 0$ for probability-turbulence divergence


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

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


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

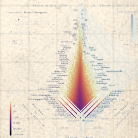


Limit of $\alpha=0$ for probability-turbulence divergence

 Normalization:



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

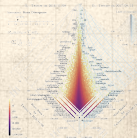
 Because the normalization also diverges as $1/\alpha$, the divergence will be zero when there are no exclusive types and non-zero when there are exclusive types.



Combine these cases into a single expression:

$$D_0^P(P_1 \parallel P_2) = \frac{1}{(N_1 + N_2)} \sum_{\tau \in R_{1,2;0}} (\delta_{p_{\tau,1},0} + \delta_{0,p_{\tau,2}}). \quad (20)$$

-  The term $(\delta_{p_{\tau,1},0} + \delta_{0,p_{\tau,2}})$ 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,

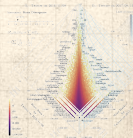


Type contribution ordering for the limit of $\alpha=0$

🧱 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 $D_0^P(P_1 \parallel P_2)$, we can still order them according to the log ratio of their probabilities.

🧱 The overall ordering of types by divergence contribution for $\alpha=0$ is then: (1) exclusive types by descending probability and then (2) types appearing in both systems by descending log ratio.

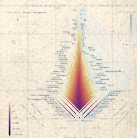


Limit of $\alpha=\infty$ for probability-turbulence divergence





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

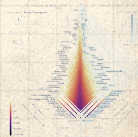
where

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



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].



Ω_1 : Twitter on 2016/11/09

Ω_2 : Twitter on 2017/08/13

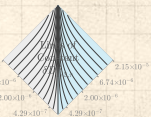
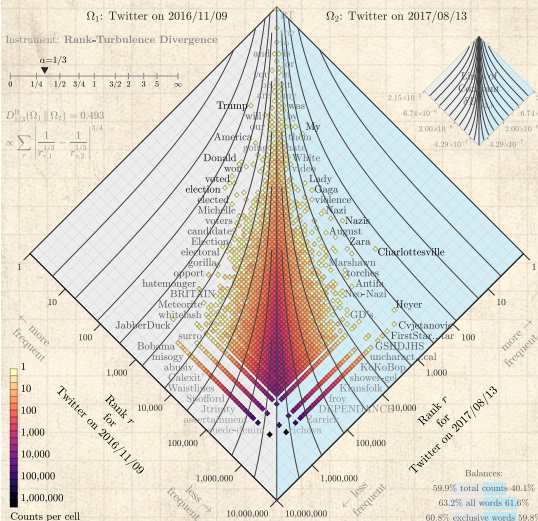
Divergence contribution $\delta D_{1/3,7}^R$ ($\times 10^{-3}\%$)

Instrument: Rank-Turbulence Divergence

$\alpha=1/3$

$$D_{1/3}^R(\Omega_1 || \Omega_2) = 0.493$$

$$\propto \sum_r \left| \frac{1}{r_{-1/2}} - \frac{1}{r_{+1/2}} \right|^{3/4}$$



Word	Count	Divergence Contribution $\delta D_{1/3,7}^R$ ($\times 10^{-3}\%$)
Trump	11	60
election	64	2,055
voted	58	1,002
Hillary	70	1,505
Donald	50	566
Nazis	48	500
president	48	500
trump	77	1,357
My	139	20
Larsson	37,952	268
supremacists	5,873	171
Zara	25,126	267
won	69	536
Heyer	862,482	443
Clinton	125	1,761
elected	151	2,787
Nazi	3,485	174
condemn	13,329	280
America	40	164
BTS	3,503	192
his	86	27
Gaga	1,175	124
Cvjetanovic	1,562,865	673
Obama	76	378
wins	144	1,209
violence	1,671	170
August	7,911	321
Lady	801	110
nazis	18,804	442
Heather	16,317	140
electoral	446	15,272
torches	47,558	610
JabberDuck	993	840,082
hatemonger	756	120,186
Antifa	102,414	743
oport	672	62,603
neo-nazis	56,244	657
Harambe	658	56,005
Michelle	261	3,115

Balances:
 59.9% total counts 40.1%
 63.2% all words 61.6%
 60.8% exclusive words 59.8%

50.2%—49.8%

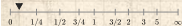
Ω_1 : Twitter on 2016/11/09

Ω_2 : Twitter on 2017/08/13

Divergence contribution $\delta D_{1/12,r}^D (\times 10^{-4}\%)$

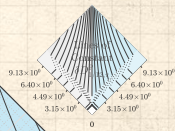
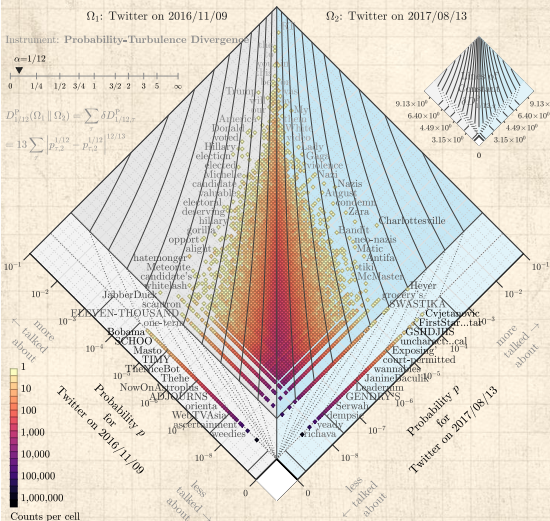
Instrument: Probability-Turbulence Divergence

$\alpha=1/12$



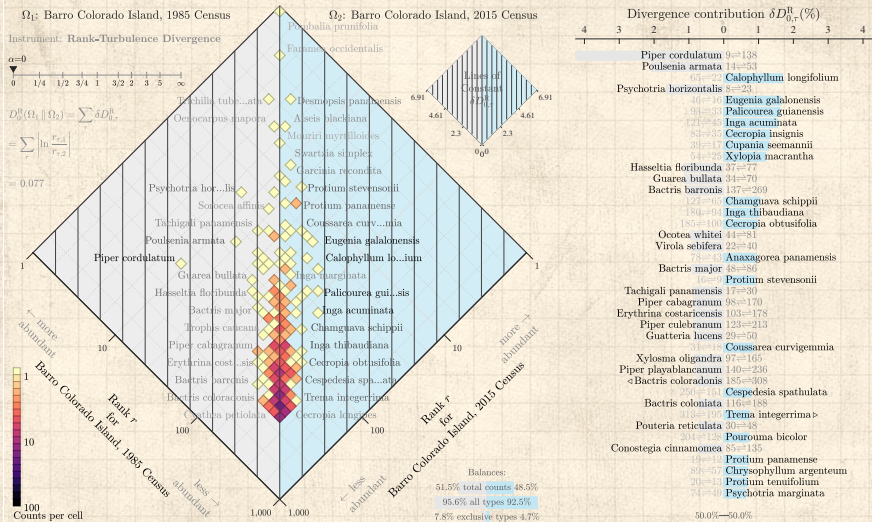
$$D_{1/12}^D(\Omega_1 \parallel \Omega_2) = \sum \delta D_{1/12,r}^D$$

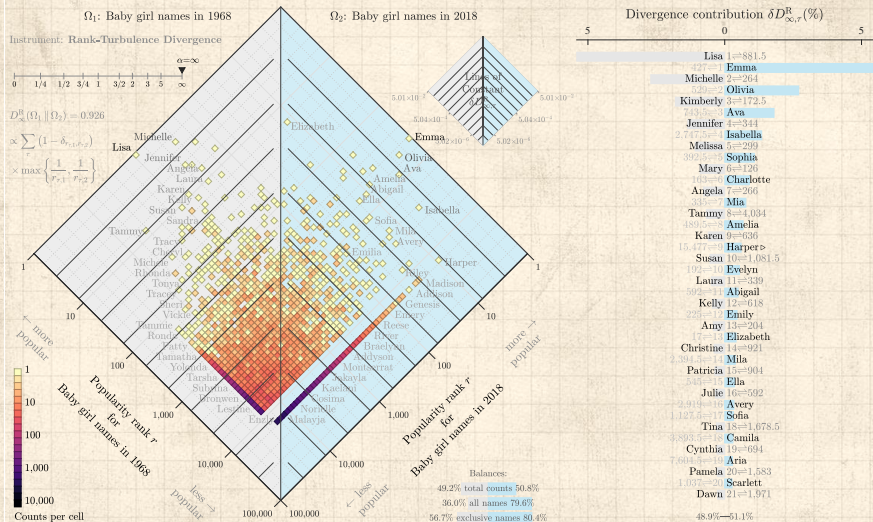
$$= 13 \sum_{P_{r,2}}^{1/12} \frac{1/12}{P_{r,2}} - \frac{1/12}{P_{r,2}^{12/13}}$$

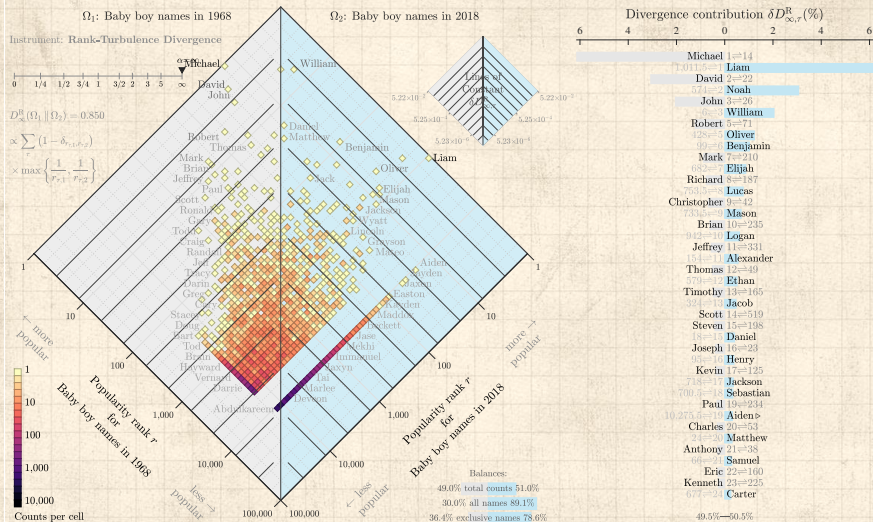


Term	Divergence contribution $\delta D_{1/12,r}^D (\times 10^{-4}\%)$
Cvijetanic >	1.552,865=6.73
FirstStarMagicAllStar >	1.552,865=1.116
KISSMARCREDDY >	1.552,865=1.47
AllStarGames >	1.552,865=1.520
Kafeel >	1.552,865=1.985
Starbz >	1.552,865=2.021
< Bobama	2.423=1,537,471
< Oarack	2.425=1,537,471
< Un-Leashed	2.703=1,537,471
GSHDJHS >	1.552,865=3.088
Bodak >	1.552,865=3.099
< KiligTripSaBagnio	3.142=1,537,471
< Somali-American	3.229=1,537,471
< DICKASS	3.321=1,537,471
< Michelle	3.412=1,537,471
Eastwatch >	1.552,865=3.673
< Un-leashed	3.645=1,537,471
Heyer's >	1.552,865=3.983
SCHOO >	3.921=1,537,471
uncharacteristical >	1.552,865=4.382
allejones >	1.552,865=4.518
misogy >	4.328=1,537,471
TLC's >	1.552,865=4.723
SORIBADA >	1.552,865=4.913
tRyNna >	4.660=1,537,471
aLmoSt >	4.671=1,537,471
tcas >	1.552,865=5.240
< Ruline	5.097=1,537,471
< Steinger	5.118=1,537,471
low-rise >	1.552,865=5.436
< climate-denying	5.191=1,537,471
CLITORIS >	1.552,865=5.662
Adityanath >	1.552,865=5.682
< lambo's	5.383=1,537,471
DelHiHasret >	1.552,865=5.755
FikBel >	1.552,865=5.757
Walker-Peters >	1.552,865=5.808
< KBAT	5.617=1,537,471
UNIDAS >	1.552,865=6.040
< stammered	5.653=1,537,471

49.9%—50.1%







Ω_1 : 1948 Google Books Fiction

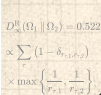
Ω_2 : 1987 Google Books Fiction

Instrument: Rank-Turbulence Divergence

$$D_{\infty}^R(\Omega_1, \Omega_2) = 0.522$$

$$\infty \sum_{\tau} (1 - \delta_{\tau,1} \delta_{\tau,2})$$

$$\times \max \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\}$$



$$D_{\infty}^R(\Omega_1, \Omega_2) = 0.522$$

$$\infty \sum_{\tau} (1 - \delta_{\tau,1} \delta_{\tau,2})$$

$$\times \max \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\}$$

$$\times \max \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\}$$

$$\times \max \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\}$$

$$\times \max \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\}$$

$$\times \max \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\}$$

$$\times \max \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\}$$

$$\times \max \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\}$$

$$\times \max \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\}$$

$$\times \max \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\}$$

$$\times \max \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\}$$

$$\times \max \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\}$$

$$\times \max \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\}$$

$$\times \max \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\}$$

$$\times \max \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\}$$

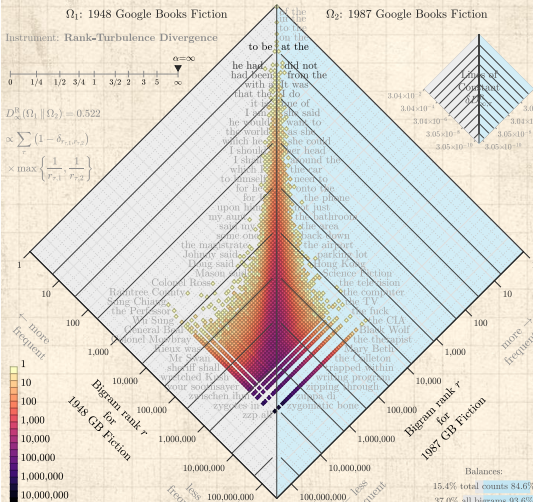
$$\times \max \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\}$$

$$\times \max \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\}$$

$$\times \max \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\}$$

$$\times \max \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\}$$

$$\times \max \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\}$$



Counts per cell

Counts per cell

Counts per cell

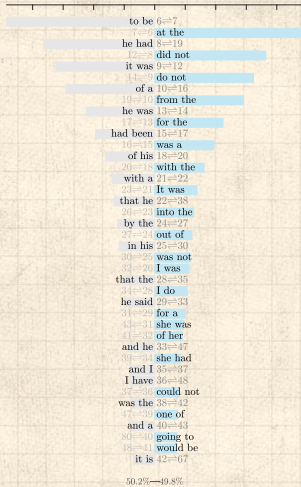
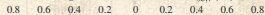
Counts per cell

Counts per cell

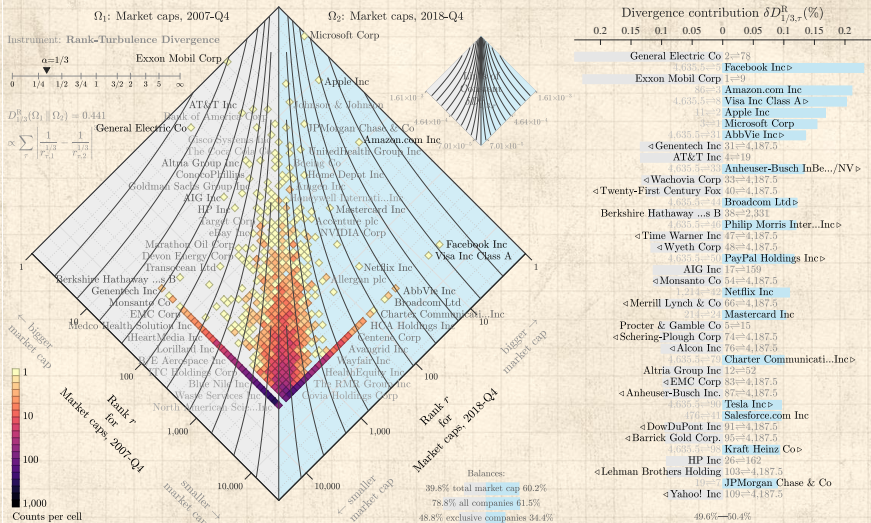
Counts per cell

Counts per cell

Divergence contribution $\delta_{\infty,r}^R$ (%)



Balances:
 15.4% total counts 84.6%
 37.0% all bigrams 93.6%
 17.2% exclusive bigrams 67.3%



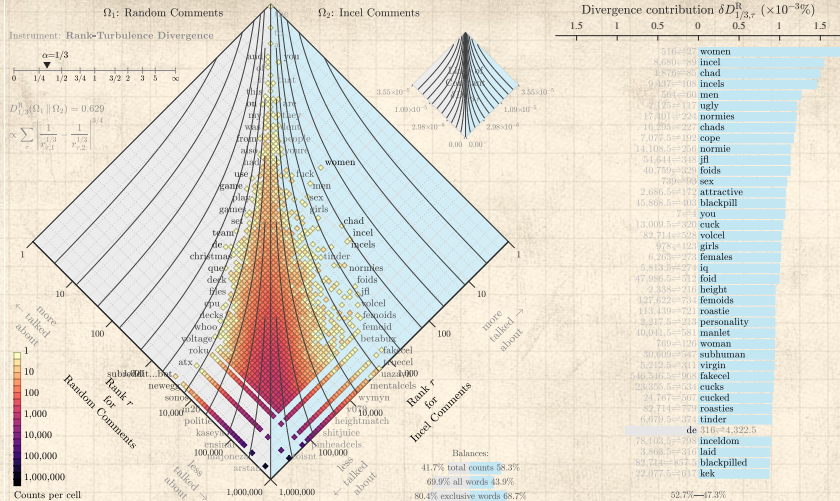
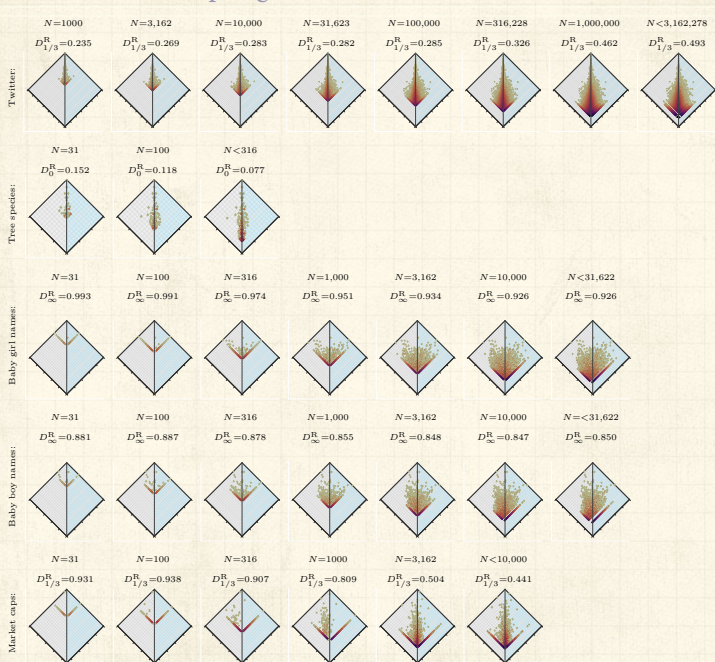


FIG. 8. Rank-turbulence divergence allotaxonograph [34] of word rank distributions in the incel vs random comment corpora. The rank-rank histogram on the left shows the density of words by their rank in the incel comments corpus against their rank in the random comments corpus. Words at the top of the 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:



The PoCverse
Allotaxonomy
54 of 73

A plentitude of
distances

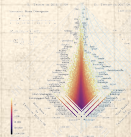
Rank-turbulence
divergence

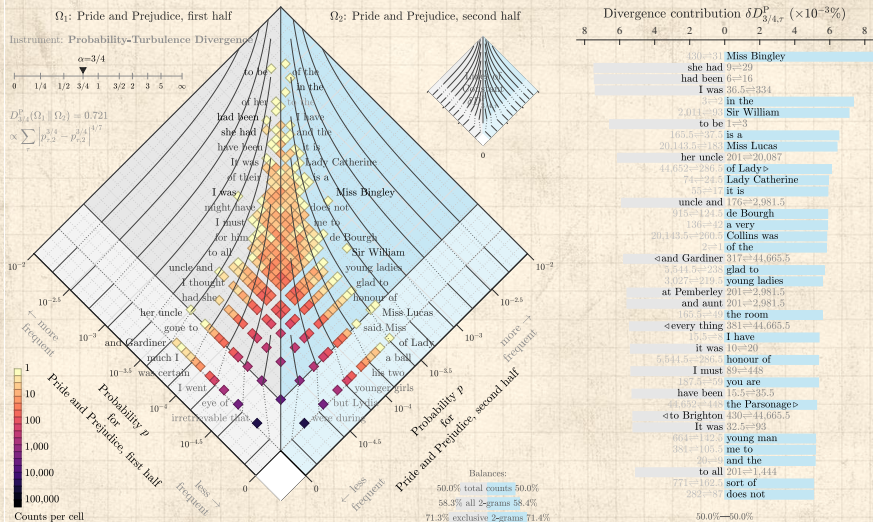
Probability-
turbulence divergence

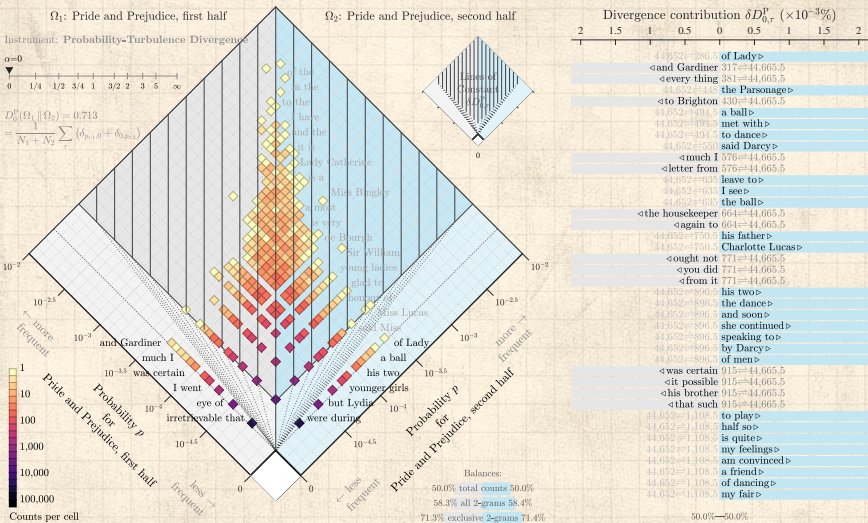
Explorations

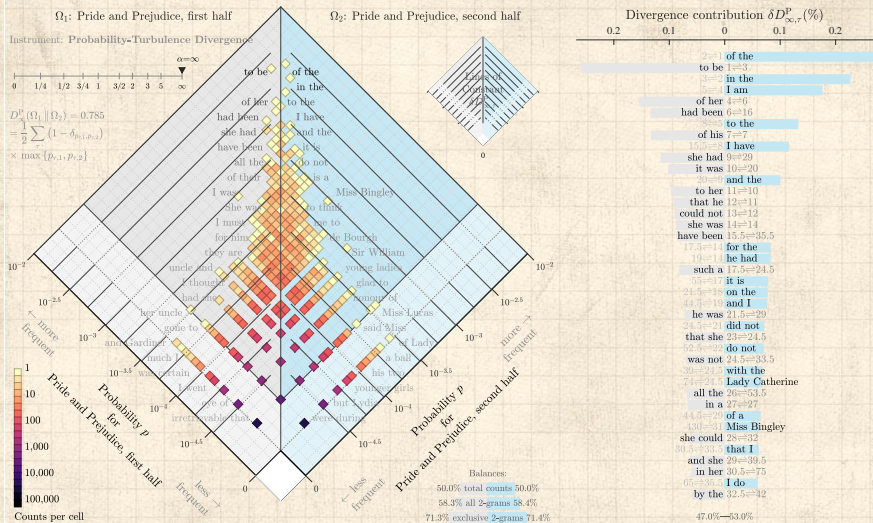
Nutshell

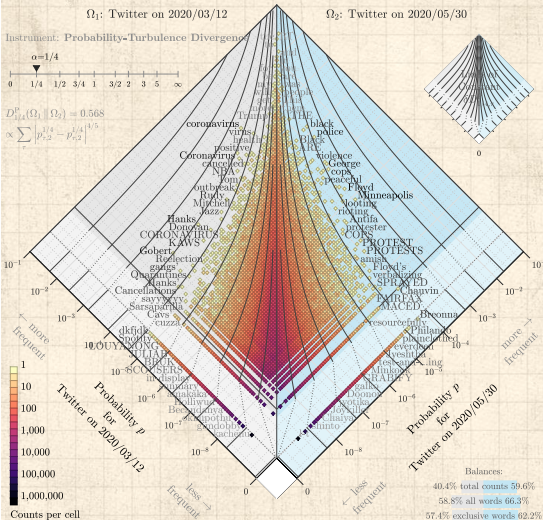
References





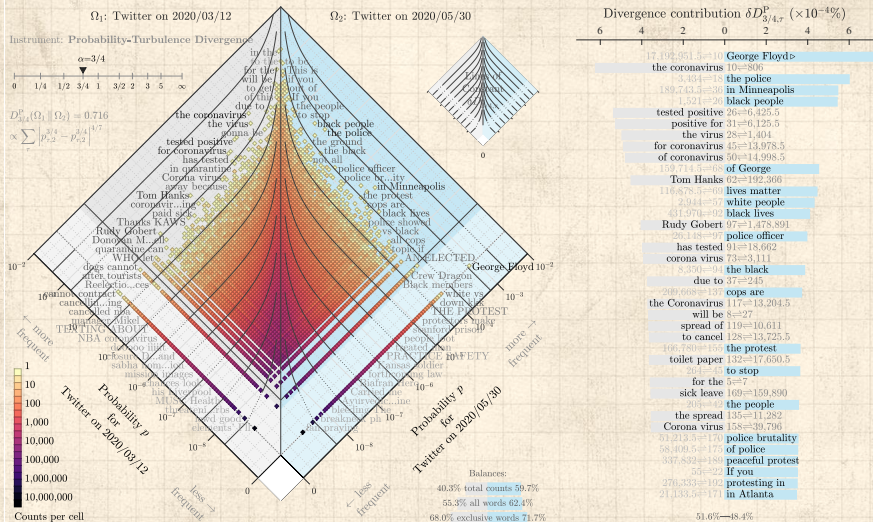






Divergence contribution $\delta D_{1/4,7}^P (\times 10^{-4}\%)$

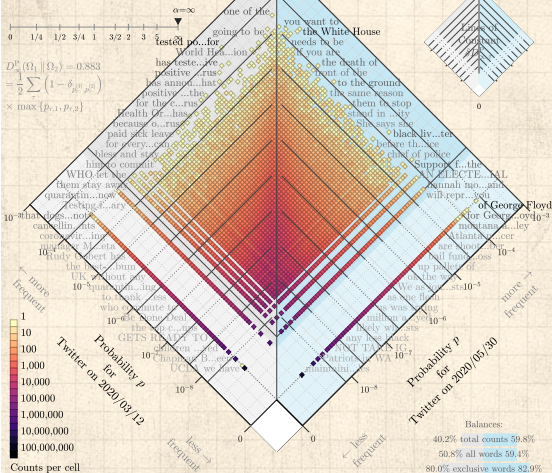
5	0	5
	29,291.5=110	Minneapolis
	17,877=117	Floyd
	coronavirus	34=708
	1,160=40	police
	4,689=8	George
	14,094.5=150	protesting
	6,661=103	cops
	4,472=87	protest
	19,096.5=194	protestors
	12,852.5=160	protesters
	22,116.5=223	looting
	Gobert	427=164,877.5
	Hanks	329=48,943.5
	611=40	black
	Coronavirus	72=2,352
	5,622=137	protests
	85,437=579	PROTEST
	913,207.5=921	Chauvin
	5,274=170	Atlanta
	43,060.5=499	Antifa
	6,784=206	peaceful
	1,717,011.5=1,327	Breonna >
	18,219=340	NYPD
	619=57	white
	tested	118=2,740
	40,175.5=520	BLM
	129,716.5=798	PROTESTS
	22,714=414	rioting
	2,797=138	violence
	4,488=180	officer
	90,270.5=77	ANTIFA
	7,925=258	cop
	NBA	198=4,648.5
	KAWS	852=65,437.5
	Rudy	408=13,205.5
	gangs	1,232=217,899.5
	46,280=76	protester
	92,973=1,02	PEACEFUL
	220,871.5=1,311	testing
		119=1,842
		Floyd's
		48.6%=51.4%



Ω_1 : Twitter on 2020/03/12

Ω_2 : Twitter on 2020/05/30

Instrument: Probability-Turbulence Divergence

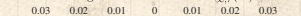


$$D_{\infty}^p(\Omega_1, \Omega_2) = 0.883$$

$$= \frac{1}{2} \sum (1 - \delta_{p, \mu_i}^{p, \mu_i})$$

$$\times \max\{p_{r,1}, p_{r,2}\}$$

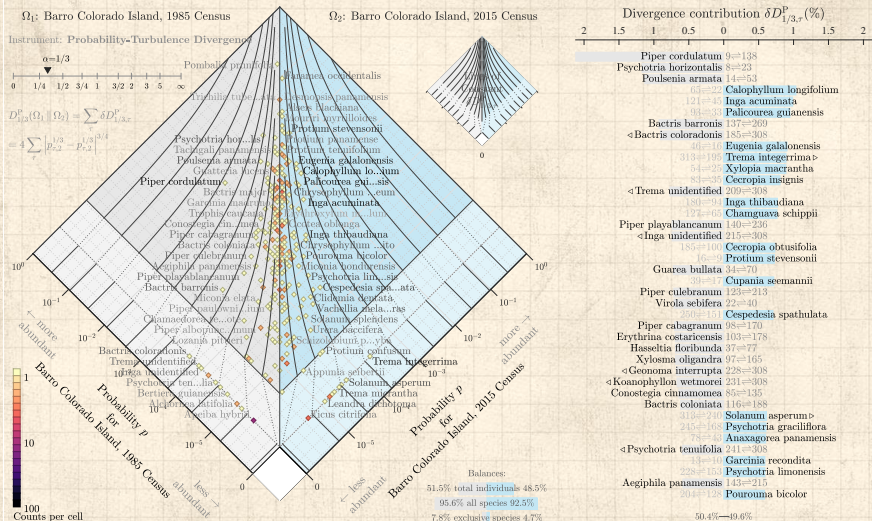
Divergence contribution $\delta D_{\infty, r}^p$ (%)



- 1=4,975.5
- 2=219.0
- 3=11,879
- 4=14,798
- 5=7,264.5
- 6=33
- 7=108
- 8=1,420
- 9=78,795
- 10=53,912
- 11=603
- 12=22,783.5
- 13=45
- 14=143.5
- 15=30
- 16=277,424.5
- 17=631.5
- 18=43,073,107
- 19=22
- 20=43,073,107
- 21=172,568
- 22=1,421
- 23=43,073,107
- 24=43,073,107
- 25=10
- 26=7
- 27=131
- 28=108
- 29=10
- 30=11
- 31=12
- 32=13
- 33=10
- 34=10
- 35=10
- 36=10
- 37=10
- 38=10
- 39=10
- 40=10
- 41=10
- 42=10
- 43=10
- 44=10
- 45=10
- 46=10
- 47=10
- 48=10
- 49=10
- 50=10
- 51=10
- 52=10
- 53=10
- 54=10
- 55=10
- 56=10
- 57=10
- 58=10
- 59=10
- 60=10
- 61=10
- 62=10
- 63=10
- 64=10
- 65=10
- 66=10
- 67=10
- 68=10
- 69=10
- 70=10
- 71=10
- 72=10
- 73=10
- 74=10
- 75=10
- 76=10
- 77=10
- 78=10
- 79=10
- 80=10

Balances:
 40.2% total counts 59.8%
 50.8% all words 59.4%
 80.0% exclusive words 82.9%

50.4%—49.6%







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](#)  

A plenitude of
distances



Rank-turbulence
divergence



Probability-
turbulence divergence



Explorations

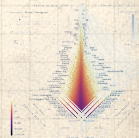


Google books:

[allotaxonometer-flipbook-7-google-books-onegrams-rank-div.pdf](#)  

[allotaxonometer-flipbook-8-google-books-bigrams-rank-div.pdf](#)   references

[allotaxonometer-flipbook-9-google-books-trigrams-rank-div.pdf](#)  







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  

A plenitude of
distances

Rank-turbulence
divergence

Probability-
turbulence divergence

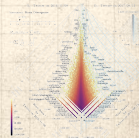
Explorations

Nutshell

References

Code:

<https://gitlab.com/compstorylab/allotaxonomer>



Claims, exaggerations, reminders:

- Needed for comparing large-scale complex systems: Comprehensible, 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:

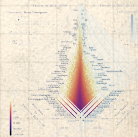
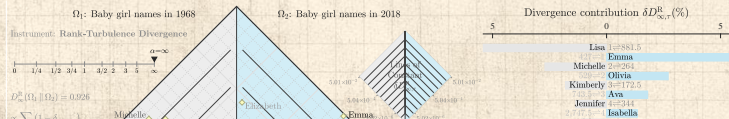
App version: <https://allotaxp.vercel.app/>

Observable version:



<https://observablehq.com/@jstonge/allotaxonometer-4-all>

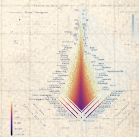
Github: <https://github.com/jstonge/allotaxp>

- Future: Probability-turbulence divergence plus many other instruments.



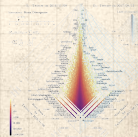
References I

- [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.
Allotaxonomy 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 allotaxonomic 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.

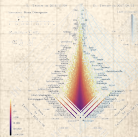
A new metric for probability distributions.

IEEE Transactions on Information theory, 2003. pdf ↗

- [8] E. Hellinger.

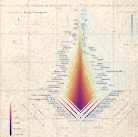
Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen.

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. Oesterreicher 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.

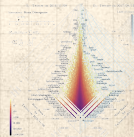
[14] E. A. Pechenick, C. M. Danforth, and P. S. Dodds.

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 

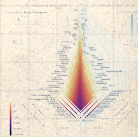
[15] Y. Sasaki.

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

The PoCverse
Allotaxonomy
73 of 73

A plenitude of
distances

Rank-turbulence
divergence

Probability-
turbulence divergence

Explorations

Nutshell

References

[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](#) 