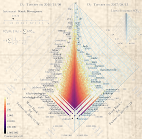
Allotaxonometry

Last updated: 2022/08/28, 03:24:52 EDT

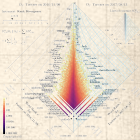
Principles of Complex Systems, Vols. 1, 2, & 3D
CSYS/MATH 300, 303, & 394, 2022–2023 | @pocsvox

Prof. Peter Sheridan Dodds | @peterdodds

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



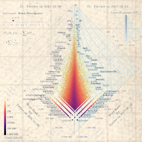
These slides are brought to you by:



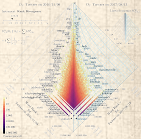
These slides are also brought to you by:

Special Guest Executive Producer

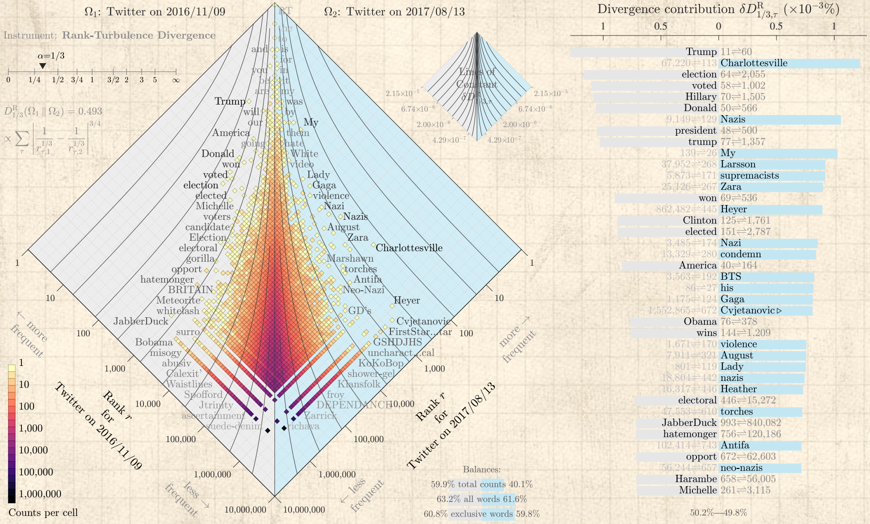
On Instagram at pratchett_the_cat



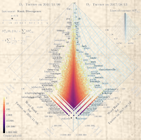
Outline



Goal—Understand this:

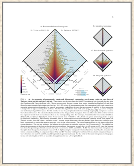


The Boggoracle Speaks:

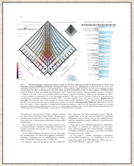


Site (papers, examples, code):
http://compstorylab.org/allotaxonometry/

Foundational papers:



"Allotaxonometry and rank-turbulence divergence: A universal instrument for comparing complex systems"
Dodds et al.,
, 2020. [5]

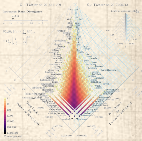


"Probability-turbulence divergence: A tunable allotaxonometric instrument for comparing heavy-tailed categorical distributions"
Dodds et al.,
, 2020. [6]

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]

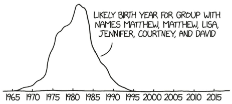
just a decade or so. If you were born in the United States around this year, these are names that are more likely to seem common and generic to *you*, but are distinctive generational markers.

1880 *Will, Maude, Minnie, May, Cora, Ida, Lula, Hattie, Jennie, Ada*
1885 *Grover, Maude, Will, Minnie, Lizzie, Effie, May, Cora, Lula, Nettie*
1890 *Maude, May, Minnie, Effie, Mabel, Bessie, Nettie, Hattie, Lula, Cora*
1895 *Maude, Mabel, Minnie, Bessie, Mamie, Myrtle, Hattie, Pearl, Ethel, Bertha*
1900 *Mabel, Myrtle, Bessie, Mamie, Pearl, Blanche, Gertrude, Ethel, Minnie, Gladys*
1905 *Gladys, Viola, Mabel, Myrtle, Gertrude, Pearl, Bessie, Blanche, Mamie, 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, June, Betty, Marjorie, Dorothy, Lorraine, Lois, Norma, Virginia, Juanita*
1930 *Dolores, Betty, Joan, Billie, Doris, Norma, Lois, Billy, June, Marilyn*
1935 *Shirley, Marlene, Joan, Dolores, Marilyn, Bobby, Betty, Billy, Joyce, Beverly*
1940 *Carole, Judith, Judy, Carol, Joyce, Barbara, Joan, Carolyn, Shirley, Jerry*
1945 *Judy, Judith, Linda, Carol, Sharon, Sandra, Carolyn, Larry, Janice, Dennis*
1950 *Linda, Deborah, Gail, Judy, Gary, Larry, Diane, Dennis, Brenda, Janice*
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, Dawn, Michele*
1970 *Tammy, Tonya, Tracy, Todd, Dawn, Tina, Stacey, Stacy, Michele, Lisa*
1975 *Chad, Jason, Tonya, Heather, Jennifer, Amy, Stacy, Shannon, Stacey, Tara*
1980 *Brandy, Crystal, April, Jason, Jeremy, Erin, Tiffany, Jamie, Melissa, Jennifer*
1985 *Krystal, Lindsay, Ashley, Lindsey, Dustin, Jessica, Amanda, Tiffany, Crystal, Amber*
1990 *Brittany, Chelsea, Kelsey, Cody, Ashley, Courtney, Kayla, Kyle, Megan, Jessica*
1995 *Taylor, Kelsey, Dakota, Austin, Haley, Cody, Tyler, Shelby, Brittany, Kayla*
2000 *Destiny, Madison, Haley, Sydney, Alexis, Kaitlyn, Hunter, Brianna, Hannah, Alyssa*
2005 *Aidan, Diego, Gavin, Hailey, Ethan, Madison, Ava, Isabella, Jayden, Aiden*
2010 *Jayden, Aiden, Nevaeh, Addison, Brayden, Landon, Peyton, Isabella, Ava, Liam*
2015 *Aria, Harper, Scarlett, Jaxon, Grayson, Lincoln, 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.

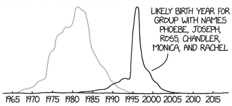
But names can reveal things about age in other ways.

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:

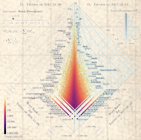


The actors were actually born in the late 1960s, on the very early edge of the popularity of their names. In other words, the actors all have names that were a little before their time. Courtney Cox and Jennifer Aniston had names that didn't really become popular until a decade later. (Maybe people with 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:

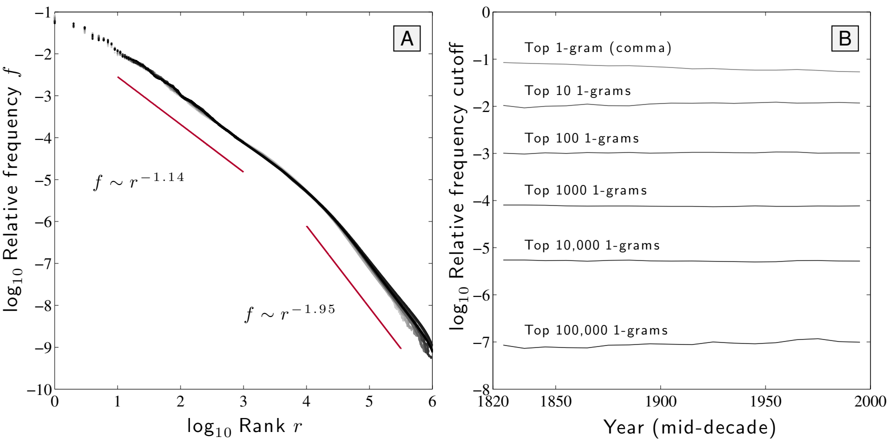


The show debuted in 1994. There's a clear spike in popularity of the names in 1995 and 1996, which can probably be attributed to the show putting the names in the minds of new parents. But it's not just the show—that name combination was clearly on the rise in the years before *Friends* premiered. It's 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?

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



For language, Zipf's law has two scaling regimes: [18]

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

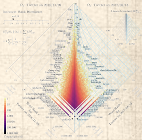
When comparing two texts, define Lexical turbulence as flux of words across a frequency threshold:

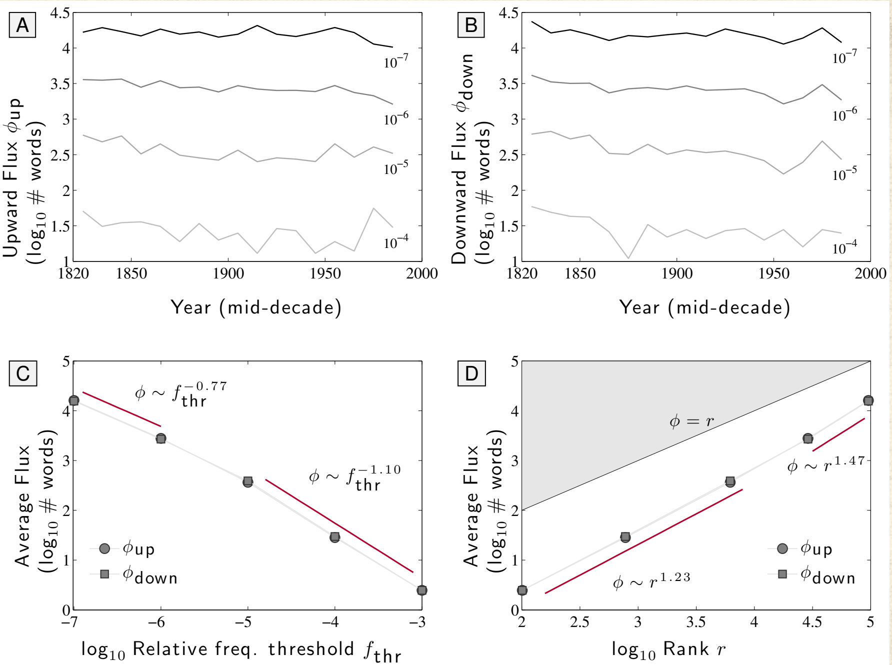
$$\phi \sim \begin{cases} f_{\text{thr}}^{-\mu} \text{ for } f_{\text{thr}} \ll f_{\text{b}}, \\ f_{\text{thr}}^{-\mu'} \text{ for } f_{\text{thr}} \gg f_{\text{b}}, \end{cases}$$

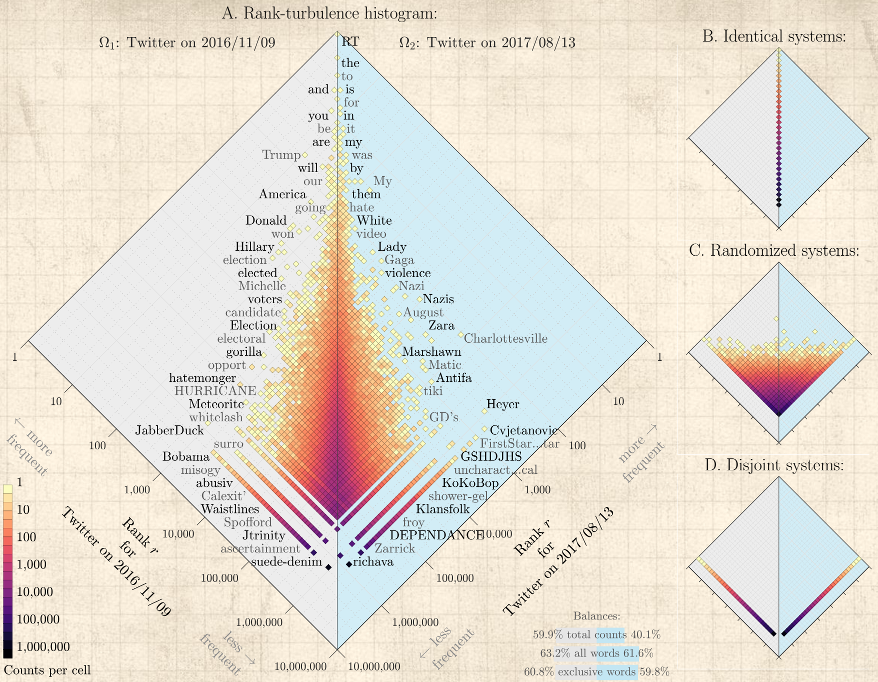
Estimates: $\mu \simeq 0.77$ and $\mu' \simeq 1.10$, and f_{b} is the scaling break point.

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

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

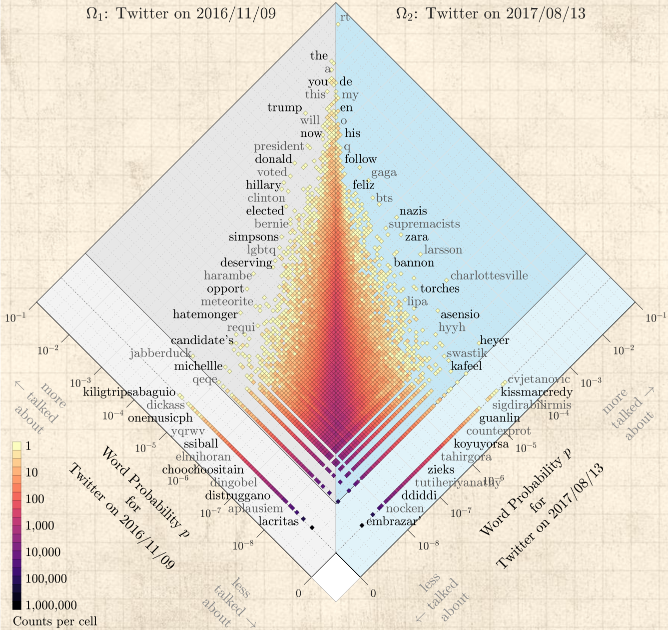


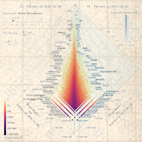

A
Upward Flux ϕup
($\log_{10}$ # words)
Year (mid-decade)
10^{-7}
10^{-6}
10^{-5}
10^{-4}
B
Downward Flux ϕdown
($\log_{10}$ # words)
Year (mid-decade)
C
Average Flux
($\log_{10}$ # words)
$\phi \sim f_{\rm thr}^{-0.77}$
$\phi \sim f_{\rm thr}^{-1.10}$
$\phi_{\rm up}$
$\phi_{\rm down}$
$\log_{10}$ Relative freq. threshold $f_{\rm thr}$
D
$\phi = r$
$\phi \sim r^{1.47}$
$\phi \sim r^{1.23}$
$\log_{10}$ Rank r



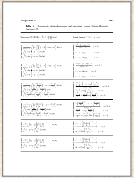
A. Rank-turbulence histogram:
Ω_1: Twitter on 2016/11/09
Ω_2: Twitter on 2017/08/13
RT
the
to
and
is
for
you
in
be
it
are
my
Trump
was
will
by
our
My
America
them
going
hate
Donald
White
won
video
Hillary
Lady
election
Gaga
elected
violence
Michelle
Nazi
voters
Nazis
candidate
August
Election
Zara
electoral
Charlottesville
gorilla
Marshawn
opport
Matic
hatemonger
Antifa
HURRICANE
tiki
Meteorite
Heyer
whitelash
GD's
JabberDuck
Cvjetanovic
surro
FirstStar...tar
Bobama
GSHDJHS
misogy
uncharact...cal
abusiv
KoKoBop
Calexit'
shower-gel
Waistlines
Klansfolk
Spofford
froy
Jtrinity
DEPENDANCE
ascertainment
Zarrick
suede-denim
richava
1
10
100
1,000
10,000
100,000
1,000,000
10,000,000
← more frequent
more frequent →
less frequent →
← less frequent
Rank r for Twitter on 2016/11/09
Rank r for Twitter on 2017/08/13
1
10
100
1,000
10,000
100,000
1,000,000
Counts per cell
Balances:
59.9% total counts 40.1%
63.2% all words 61.6%
60.8% exclusive words 59.8%
B. Identical systems:
C. Randomized systems:
D. Disjoint systems:

Zipf-turbulence histogram for probability:

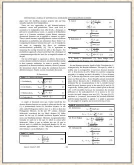




So, so many ways to compare probability distributions:

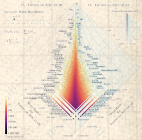


"Families of Alpha- Beta- and Gamma- Divergences: Flexible and Robust Measures of Similarities"
Cichocki and Amari,
Entropy, **12**, 1532-1568, 2010. [2]



"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 ...
- A worry: Subsampled distributions with very heavy tails
- 60ish kinds of comparisons grouped into 10 families



Quite the festival:

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

Table 3. Intersection family		
11. Intersection	$s_{IS} = \sum_{i=1}^{d} \min(P_i, Q_i)$	(12)
	$d_{non-IS} = 1 - s_{IS} = \frac{1}{2} \sum_{i=1}^{d} \lvert P_i - Q_i \rvert$	(13)
12. Wave Hedges	$d_{WH} = \sum_{i=1}^{d} (1 - \frac{\min(P_i, Q_i)}{\max(P_i, Q_i)})$	(14)
	$= \sum_{i=1}^{d} \frac{\lvert P_i - Q_i \rvert}{\max(P_i, Q_i)}$	(15)
13. Czekanowski	$s_{Cze} = \frac{2 \sum_{i=1}^{d} \min(P_i, Q_i)}{\sum_{i=1}^{d} (P_i + Q_i)}$	(16)
	$d_{Cze} = 1 - s_{Cze} = \frac{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert}{\sum_{i=1}^{d} (P_i + Q_i)}$	(17)
14. Motyka	$s_{Mot} = \frac{\sum_{i=1}^{d} \min(P_i, Q_i)}{\sum_{i=1}^{d} (P_i + Q_i)}$	(18)
	$d_{Mot} = 1 - s_{Mot} = \frac{\sum_{i=1}^{d} \max(P_i, Q_i)}{\sum_{i=1}^{d} (P_i + Q_i)}$	(19)
15. Kulczynski s	$s_{Kul} = \frac{1}{d_{Kul}} = \frac{\sum_{i=1}^{d} \min(P_i, Q_i)}{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert}$	(20)
16. Ruzicka	$s_{Ruz} = \frac{\sum_{i=1}^{d} \min(P_i, Q_i)}{\sum_{i=1}^{d} \max(P_i, Q_i)}$	(21)
17. Tanimoto	$d_{Tani} = \frac{\sum_{i=1}^{d} P_i + \sum_{i=1}^{d} Q_i - 2\sum_{i=1}^{d} \min(P_i, Q_i)}{\sum_{i=1}^{d} P_i + \sum_{i=1}^{d} Q_i - \sum_{i=1}^{d} \min(P_i, Q_i)}$	(22)
	$= \frac{\sum_{i=1}^{d} (\max(P_i, Q_i) - \min(P_i, Q_i))}{\sum_{i=1}^{d} \max(P_i, Q_i)}$	(23)

Table 4. Inner Product family		
18. Inner Product	$s_{IP} = P \bullet Q = \sum_{j=1}^{d} P_i Q_i$	(24)
19. Harmonic mean	$s_{HM} = 2 \sum_{i=1}^{d} \frac{P_i Q_i}{P_i + Q_i}$	(25)
20. Cosine	$s_{Cos} = \frac{\sum_{i=1}^{d} P_i Q_i}{\sqrt{\sum_{i=1}^{d} P_i^2} \sqrt{\sum_{i=1}^{d} Q_i^2}}$	(26)
21. Kumar-Hassebrook (PCE)	$s_{Jac} = \frac{\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i}$	(27)
22. Jaccard	$s_{Jac} = \frac{\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i}$	(28)
	$d_{Jac} = 1 - s_{Jac} = \frac{\sum_{i=1}^{d} (P_i - Q_i)^2}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i}$	(39)
23. Dice	$s_{Dice} = \frac{2 \sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2}$	(40)
	$d_{Dice} = 1 - s_{Dice} = \frac{\sum_{i=1}^{d} (P_i - Q_i)^2}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2}$	(31)

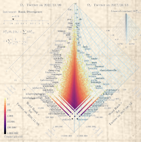
Table 5. Fidelity family or Squared-chord family		
24. Fidelity	$s_{Fid} = \sum_{i=1}^{d} \sqrt{P_i Q_i}$	(32)
25. Bhattacharyya	$d_B = -\ln \sum_{i=1}^{d} \sqrt{P_i Q_i}$	(33)
26. Hellinger	$d_H = \sqrt{2 \sum_{i=1}^{d} (\sqrt{P_i} - \sqrt{Q_i})^2}$	(34)
	$= 2\sqrt{1 - \sum_{i=1}^{d} \sqrt{P_i Q_i}}$	(35)
27. Matusita	$d_M = \sqrt{\sum_{i=1}^{d} (\sqrt{P_i} - \sqrt{Q_i})^2}$	(36)
	$= \sqrt{2 - 2 \sum_{i=1}^{d} \sqrt{P_i Q_i}}$	(37)
28. Squared-chord	$d_{sqc} = \sum_{i=1}^{d} (\sqrt{P_i} - \sqrt{Q_i})^2$	(38)
$s_{sqc} = 1 - d_{sqc}$	$s_{sqc} = 2 \sum_{i=1}^{d} \sqrt{P_i Q_i} - 1$	(39)

Table 6. Squared L_2 family or χ^2 family		
29. Squared Euclidean	$d_{sqe} = \sum_{i=1}^{d} (P_i - Q_i)^2$	(40)
30. Pearson χ^2	$d_P(P, Q) = \sum_{i=1}^{d} \frac{(P_i - Q_i)^2}{Q_i}$	(41)
31. Neyman χ^2	$d_N(P, Q) = \sum_{i=1}^{d} \frac{(P_i - Q_i)^2}{P_i}$	(42)
32. Squared χ^2	$d_{SqChi} = \sum_{i=1}^{d} \frac{(P_i - Q_i)^2}{P_i + Q_i}$	(43)
33. Probabilistic Symmetric χ^2	$d_{PChi} = 2 \sum_{i=1}^{d} \frac{(P_i - Q_i)^2}{P_i + Q_i}$	(44)
34. Divergence	$d_{Div} = 2 \sum_{i=1}^{d} \frac{(P_i - Q_i)^2}{(P_i + Q_i)^2}$	(45)
35. Clark	$d_{Clk} = \sqrt{\sum_{i=1}^{d} \left(\frac{\lvert P_i - Q_i \rvert}{P_i + Q_i} \right)^2}$	(46)
36. Additive Symmetric χ^2	$d_{AdChi} = \sum_{i=1}^{b} \frac{(P_i - Q_i)^2 (P_i + Q_i)}{P_i Q_i}$	(47)
* Squared L_2 family $\supset$ {Jaccard (29), Dice (31)}		

Table 7. Shannon's entropy family		
37. Kullback–Leibler	$d_{KL} = \sum_{i=1}^{d} P_i \ln \frac{P_i}{Q_i}$	(48)
38. Jeffreys	$d_J = \sum_{i=1}^{d} (P_i - Q_i) \ln \frac{P_i}{Q_i}$	(49)
39. K divergence	$d_{Kdiv} = \sum_{i=1}^{d} P_i \ln \frac{2P_i}{P_i + Q_i}$	(50)
40. Topsøe	$d_{Top} = \sum_{i=1}^{d} \left(P_i \ln \left(\frac{2P_i}{P_i + Q_i} \right) + Q_i \ln \left(\frac{2Q_i}{P_i + Q_i} \right) \right)$	(51)
41. Jensen-Shannon	$d_{JS} = \frac{1}{2} \left[\sum_{i=1}^{d} P_i \ln \left(\frac{2P_i}{P_i + Q_i} \right) + \sum_{i=1}^{d} Q_i \ln \left(\frac{2Q_i}{P_i + Q_i} \right) \right]$	(52)
42. Jensen difference	$d_{JD} = \sum_{i=1}^{b} \left[\frac{P_i \ln P_i + Q_i \ln Q_i}{2} - \left(\frac{P_i + Q_i}{2} \right) \ln \left(\frac{P_i + Q_i}{2} \right) \right]$	(53)

Table 8. Combinations		
43. Taneja	$d_{TJ} = \sum_{i=1}^{d} \left(\frac{P_i + Q_i}{2} \right) \ln \left(\frac{P_i + Q_i}{2\sqrt{P_i Q_i}} \right)$	(54)
44. Kumar-Johnson	$d_{KJ} = \sum_{i=1}^{d} \left(\frac{(P_i^2 - Q_i^2)^2}{2(P_i Q_i)^{3/2}} \right)$	(55)
45. Avg(L_1, L_∞)	$d_{ACC} = \frac{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert + \max_i \lvert P_i - Q_i \rvert}{2}$	(56)

Table 10. Vicissitude		
Vicis-Wave Hedges	$d_{emanon1} = \sum_{i=1}^{d} \frac{\lvert P_i - Q_i \rvert}{\min(P_i, Q_i)}$	(60)
Vicis-Symmetric χ^2	$d_{emanon2} = \sum_{i=1}^{d} \frac{(P_i - Q_i)^2}{\min(P_i, Q_i)^2}$	(61)
Vicis-Symmetric χ^2	$d_{emanon3} = \sum_{i=1}^{d} \frac{(P_i - Q_i)^2}{\min(P_i, Q_i)}$	(62)
Vicis-Symmetric χ^2	$d_{emanon4} = \sum_{i=1}^{d} \frac{(P_i - Q_i)^2}{\max(P_i, Q_i)}$	(63)
max-Symmetric χ^2	$d_{e5} = \max \left(\sum_{i=1}^{d} \frac{(P_i - Q_i)^2}{P_i}, \sum_{i=1}^{d} \frac{(P_i - Q_i)^2}{Q_i} \right)$	(64)
min-symmetric χ^2	$d_{e6} = \min \left(\sum_{i=1}^{d} \frac{(P_i - Q_i)^2}{P_i}, \sum_{i=1}^{d} \frac{(P_i - Q_i)^2}{Q_i} \right)$	(65)



- 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)

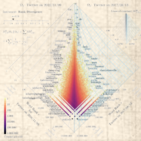
Table 1. L_p Minkowski family

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



- Information theoretic sortings are more opaque
- No tunability

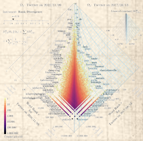
Table 1. L_p Minkowski family

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

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



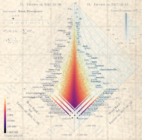
- Shannon's Entropy:

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

- Kullback-Liebler (KL) divergence:

$$D^{\mathsf{KL}}(P_2 \,||\, P_1) = \left\langle \log_2 \frac{1}{p_{2,\tau}} - \log_2 \frac{1}{p_{1,\tau}} \right\rangle_{P_2}$$
$$= \sum_{\tau \in R_{1,2;\alpha}} p_{2,\tau} \left[\log_2 \frac{1}{p_{2,\tau}} - \log_2 \frac{1}{p_{1,\tau}} \right]$$
$$= \sum_{\tau \in R_{1,2;\alpha}} p_{2,\tau} \log_2 \frac{p_{1,\tau}}{p_{2,\tau}}. \qquad (2)$$

- 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.



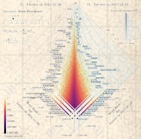
- Jensen-Shannon divergence (JSD): [9, 7, 13, 1]

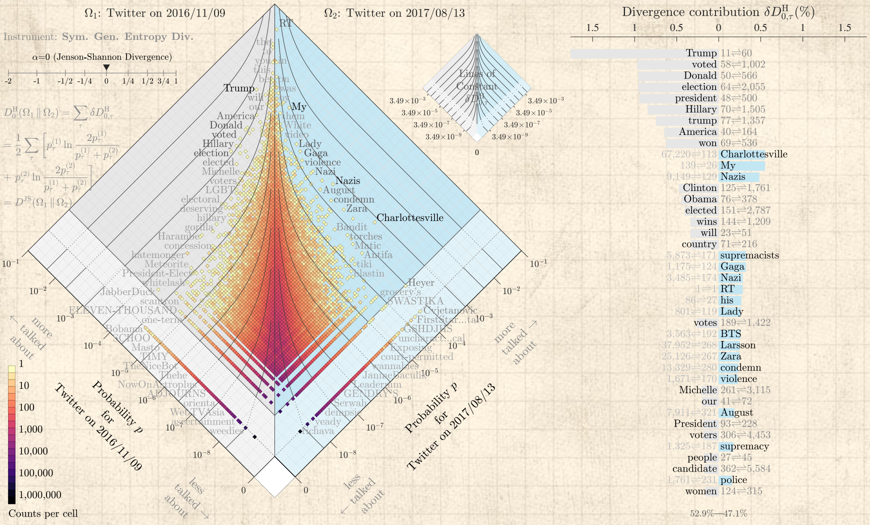
$$D^{\mathrm{JS}}\left(P_1 \,\|\, P_2\right)$$
$$= \tfrac{1}{2} D^{\mathrm{KL}}\left(P_1 \,\middle\|\, \tfrac{1}{2}\left[P_1 + P_2\right]\right) + \tfrac{1}{2} D^{\mathrm{KL}}\left(P_2 \,\middle\|\, \tfrac{1}{2}\left[P_1 + P_2\right]\right)$$
$$= \frac{1}{2} \sum_{\tau \in R_{1,2;\alpha}} \left(p_{1,\tau}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

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

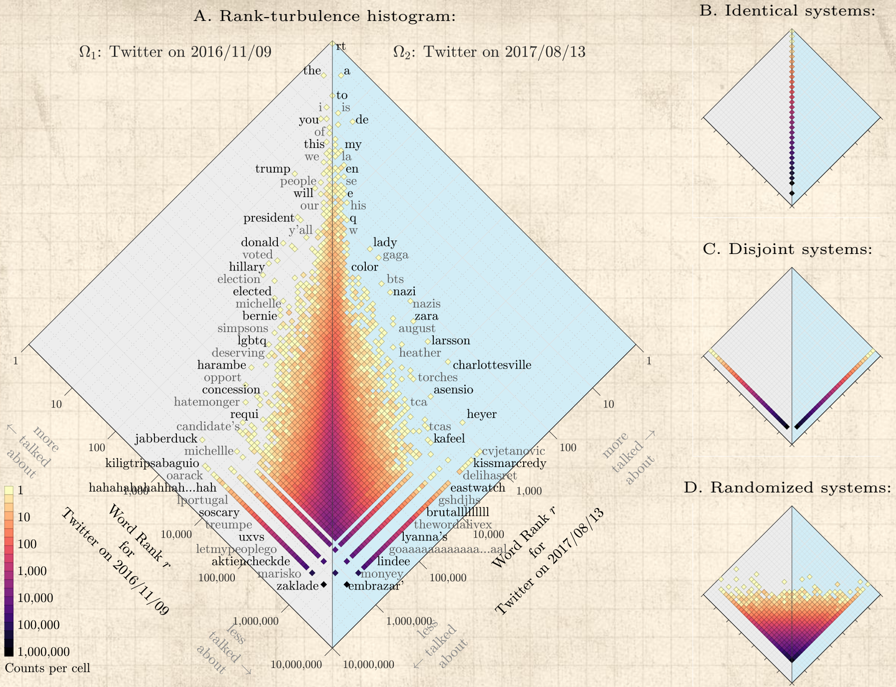
$$D^{\mathrm{AS2}}_{\alpha}\left(P_1 \,\|\, P_2\right) =$$
$$\frac{1}{\alpha(\alpha-1)} \sum_{\tau \in R_{1,2;\alpha}} \left[\left(p_{\tau,1}^{1-\alpha} + p_{\tau,2}^{1-\alpha}\right) \left(\frac{p_{\tau,1} + p_{\tau,2}}{2}\right)^{\alpha} - \left(p_{\tau,1} + p_{\tau,2}\right) \right]. \quad (4)$$

Produces JSD when $\alpha \to 0$.





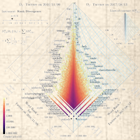
Ω1: Twitter on 2016/11/09
Ω2: Twitter on 2017/08/13
Instrument: Sym. Gen. Entropy Div.
α=0 (Jenson-Shannon Divergence)
-2 -1 -1/2 -1/4 0 1/4 1/2 3/4 1
$D_0^{H}(\Omega_1 \| \Omega_2) = \sum_\tau \delta D_{0,\tau}^{H}$
$= \frac{1}{2} \sum_\tau \left[p_\tau^{(1)} \ln \frac{2p_\tau^{(1)}}{p_\tau^{(1)} + p_\tau^{(2)}} + p_\tau^{(2)} \ln \frac{2p_\tau^{(2)}}{p_\tau^{(1)} + p_\tau^{(2)}} \right]$
$= D^{JS}(\Omega_1 \| \Omega_2)$
Lines of Constant $\delta D_{0,\tau}^{H}$
3.49×10^{-3}
3.49×10^{-5}
3.49×10^{-7}
3.49×10^{-9}
Probability p for Twitter on 2016/11/09
Probability p for Twitter on 2017/08/13
← more talked about
more talked about →
less talked about →
← less talked about
Counts per cell
1
10
100
1,000
10,000
100,000
1,000,000
Divergence contribution $\delta D_{0,\tau}^{H}$ (%)
1.5 1 0.5 0 0.5 1 1.5
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 Charlottesville
139⇌26 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 189⇌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%



A. Rank-turbulence histogram:
Ω1: Twitter on 2016/11/09
Ω2: Twitter on 2017/08/13
B. Identical systems:
C. Disjoint systems:
D. Randomized systems:
rt
the
a
to
i
is
you
de
of
this
my
we
la
trump
en
people
se
will
e
our
his
president
q
y'all
w
donald
lady
voted
gaga
hillary
color
election
bts
elected
nazi
michelle
nazis
bernie
zara
simpsons
august
lgbtq
larsson
deserving
heather
harambe
charlottesville
opport
torches
concession
asensio
hatemonger
tca
requi
heyer
candidate's
tcas
jabberduck
kafeel
michellle
cvjetanovic
kiligtripsabaguio
kissmarcredy
oarack
delihasret
hahahahahahhah...hah
eastwatch
lportugal
gshdjhs
soscary
brutallllllll
treumpe
thewordalivex
uxvs
lyanna's
letmypeoplego
goaaaaaaaaaaaa...aal
aktiencheckde
lindee
marisko
monyey
zaklade
embrazar'
1
10
100
1,000
10,000
100,000
1,000,000
10,000,000
← more talked about
more talked about →
less talked about →
← less talked about
Word Rank r for Twitter on 2016/11/09
Word Rank r for Twitter on 2017/08/13
Counts per cell

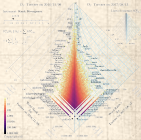
Exclusive types:

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



Desirable rank-turbulence divergence features:

1. Rank-based.
2. Symmetric.
3. Semi-positive: $D^{\mathrm{R}}_{\alpha}(\Omega_1 \,||\, \Omega_2) \geq 0$.
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. Zipfophilic: Able to handle 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’



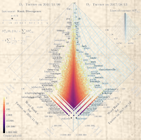
Some good things about ranks:

- 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:

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

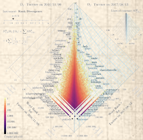
- Inverse of rank gives an increasing measure of 'importance'
- High rank means closer to rank 1
- We assign tied ranks for components of equal 'size'
- Issue: Biases toward high rank components



We introduce a tuning parameter:

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

- 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.



Trouble:

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

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

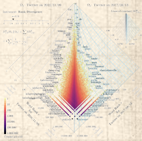
- The leading order term is:

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

  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.



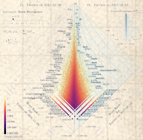
Some reworking:

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

- 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.

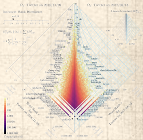
Rank-turbulence divergence:

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



Normalization:

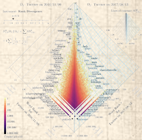
- Take a data-driven rather than analytic approach to determining $\mathcal{N}_{1,2;\alpha}$.
- Compute $\mathcal{N}_{1,2;\alpha}$ by taking the two systems to be disjoint while maintaining their underlying Zipf distributions.
- Ensures: $0 \leq D^{\mathsf{R}}_{\alpha}(R_1 \,\|\, R_2) \leq 1$
- Limits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

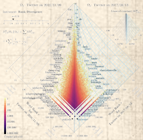
$$D^{\mathsf{R}}_{\alpha}(R_1 \,||\, R_2) = \frac{1}{\mathcal{N}_{1,2;\alpha}} \frac{\alpha+1}{\alpha} \sum_{\tau \in R_{1,2;\alpha}} \left| \frac{1}{\left[r_{\tau,1}\right]^{\alpha}} - \frac{1}{\left[r_{\tau,2}\right]^{\alpha}} \right|^{1/(\alpha+1)} . \tag{10}$$



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:

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



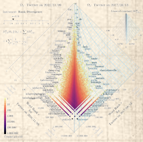
Limit of $\alpha \to 0$:

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

where

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

- Largest rank ratios dominate.



Limit of $\alpha \to \infty$:

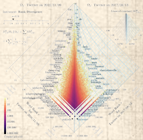
$$D^{\mathsf{R}}_{\infty}(R_1 \,\|\, R_2) = \sum_{\tau \in R_{1,2;\alpha}} \delta D^{\mathsf{R}}_{\infty,\tau}$$

$$= \frac{1}{\mathcal{N}_{1,2;\infty}} \sum_{\tau \in R_{1,2;\alpha}} \left(1 - \delta_{r_{\tau,1} r_{\tau,2}}\right) \max_{\tau} \left\{ \frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}} \right\}. \quad (14)$$

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:

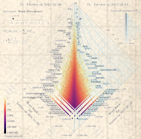
$$D^{\mathsf{P}}_{\alpha}(P_1 \,||\, P_2) = \frac{1}{\mathcal{N}^{\mathsf{P}}_{1,2;\alpha}} \frac{\alpha+1}{\alpha} \sum_{\tau \in R_{1,2;\alpha}} \left| \left[p_{\tau,1}\right]^{\alpha} - \left[p_{\tau,2}\right]^{\alpha} \right|^{1/(\alpha+1)}. \quad (16)$$

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

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

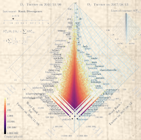


Limit of α=0 for probability-turbulence divergence

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

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

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

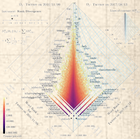


Limit of α=0 for probability-turbulence divergence

- Normalization:

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

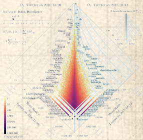
- Because the normalization also diverges as $1/\alpha$, the divergence will be zero when there are no exclusive types and non-zero when there are exclusive types.



Combine these cases into a single expression:

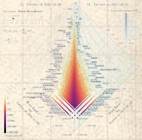
$$D_0^{\mathsf{P}}(P_1 \,\|\, P_2) = \frac{1}{(N_1 + N_2)} \sum_{\tau \in R_{1,2;0}} \left(\delta_{p_{\tau,1},0} + \delta_{0,p_{\tau,2}} \right). \quad (20)$$

- 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 $\alpha \to 0$, 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 $D_0^{\mathsf{P}}(P_1 \,\|\, P_2)$, 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.

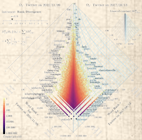


Limit of α=∞ for probability-turbulence divergence

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

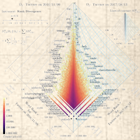
where

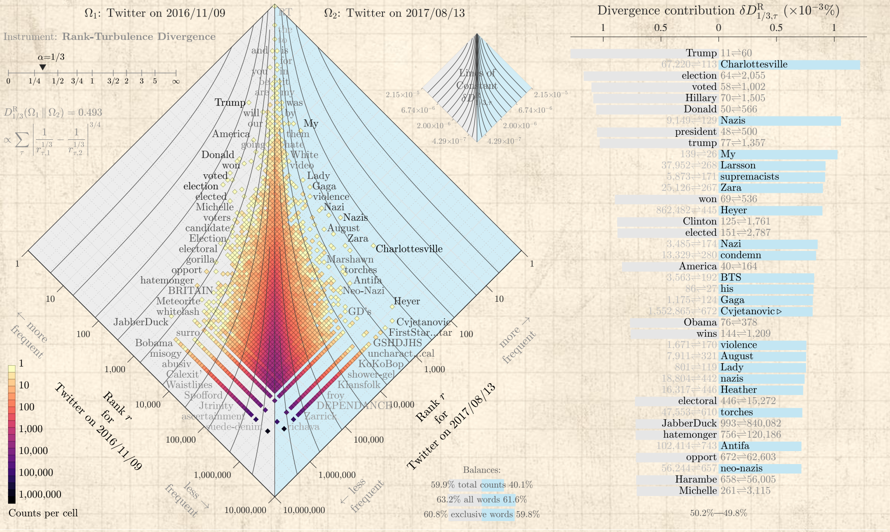
$$\mathcal{N}^{\mathsf{P}}_{1,2;\infty} = \sum_{\tau \in R_{1,2;\infty}} \left(p_{\tau,1} + p_{\tau,2} \right) = 1 + 1 = 2. \quad (22)$$



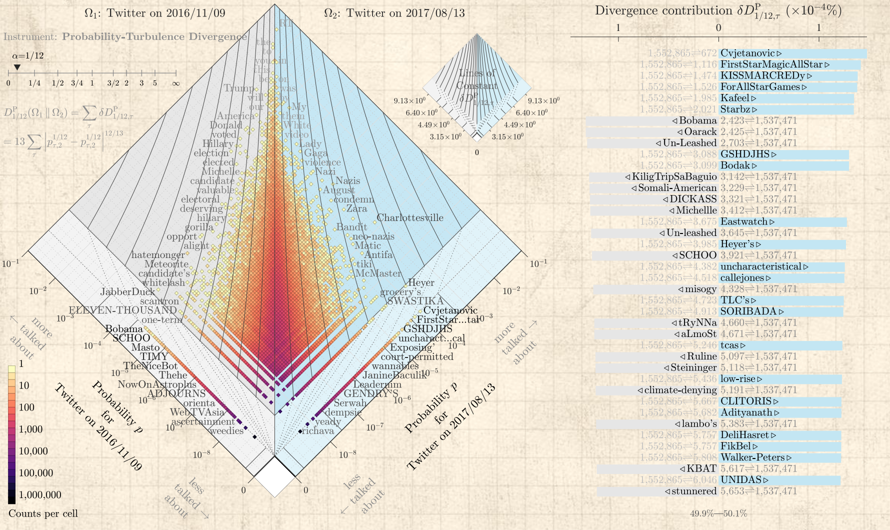
Connections for PTD:

- $\alpha = 0$: Similarity measure Sørensen-Dice coefficient [4, 16, 10], F_1 score of a test's accuracy [17, 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].

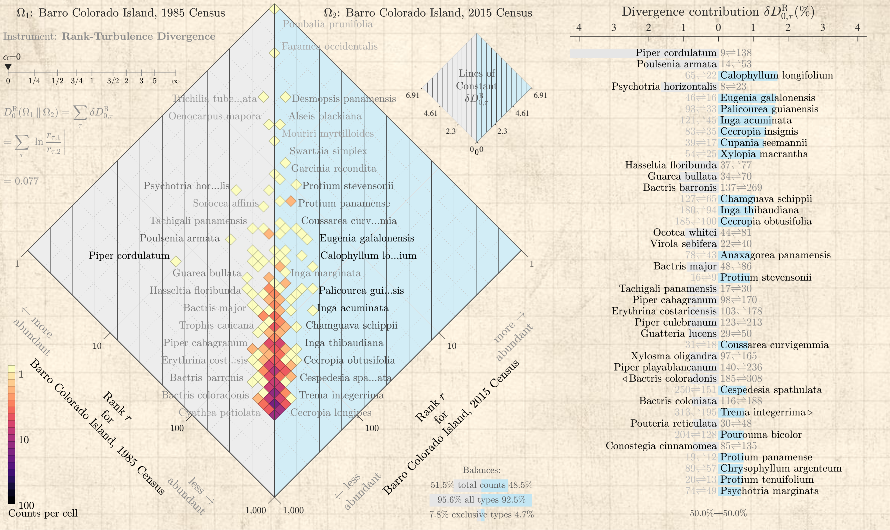




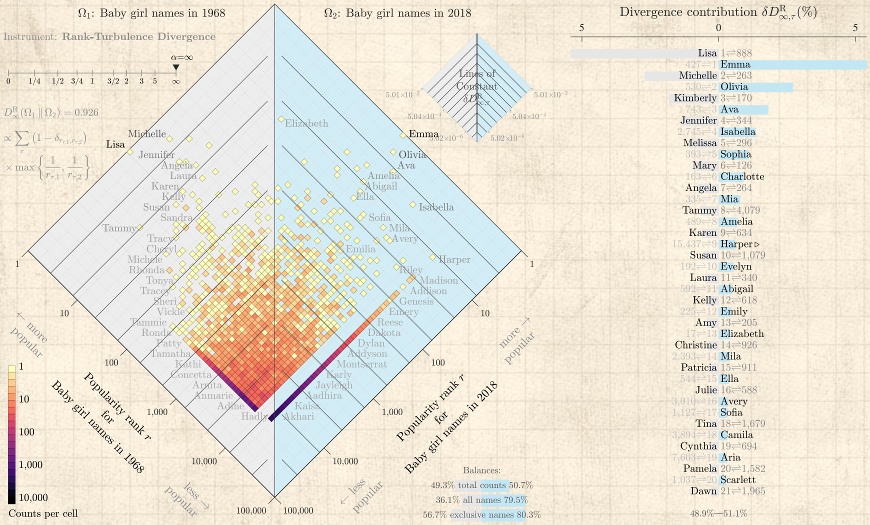
Ω1: Twitter on 2016/11/09
Ω2: Twitter on 2017/08/13
Instrument: Rank-Turbulence Divergence
α=1/3
0 1/4 1/2 3/4 1 3/2 2 3 5 ∞
$D^{R}_{1/3}(\Omega_1 \| \Omega_2) = 0.493$
$\propto \sum_{\tau} \left| \frac{1}{r_{\tau,1}^{1/3}} - \frac{1}{r_{\tau,2}^{1/3}} \right|^{3/4}$
Lines of Constant $\delta D^{R}_{1/3,\tau}$
2.15×10⁻⁵
6.74×10⁻⁶
2.00×10⁻⁶
4.29×10⁻⁷
RT
the
to
and
is
for
you
in
be
it
are
my
was
Trump
will
by
our
My
America
them
going
hate
Donald
White
won
video
voted
Lady
election
Gaga
elected
violence
Michelle
Nazi
voters
Nazis
candidate
August
Election
Zara
electoral
Charlottesville
gorilla
Marshawn
opport
torches
hatemonger
Antifa
BRITAIN
Neo-Nazi
Meteorite
Heyer
whitelash
GD's
JabberDuck
Cvjetanovic
surro
FirstStar...tar
Bobama
GSHDJHS
misogy
uncharact...cal
abusiv
KoKoBop
Calexit'
shower-gel
Waistlines
Klansfolk
Spofford
froy
Jtrinity
DEPENDANCE
ascertainment
Zarrick
suede-denim
richava
1
10
100
1,000
10,000
100,000
1,000,000
10,000,000
← more frequent
more frequent →
less frequent →
← less frequent
Rank r for Twitter on 2016/11/09
Rank r for Twitter on 2017/08/13
1
10
100
1,000
10,000
100,000
1,000,000
Counts per cell
Balances:
59.9% total counts 40.1%
63.2% all words 61.6%
60.8% exclusive words 59.8%
Divergence contribution $\delta D^{R}_{1/3,\tau}$ (×10⁻³%)
1 0.5 0 0.5 1
Trump 11⇌60
67,220⇌113 Charlottesville
election 64⇌2,055
voted 58⇌1,002
Hillary 70⇌1,505
Donald 50⇌566
9,149⇌129 Nazis
president 48⇌500
trump 77⇌1,357
139⇌26 My
37,952⇌268 Larsson
5,873⇌171 supremacists
25,126⇌267 Zara
won 69⇌536
862,482⇌445 Heyer
Clinton 125⇌1,761
elected 151⇌2,787
3,485⇌174 Nazi
13,329⇌280 condemn
America 40⇌164
3,563⇌192 BTS
86⇌27 his
1,175⇌124 Gaga
1,552,865⇌672 Cvjetanovic ▷
Obama 76⇌378
wins 144⇌1,209
1,671⇌170 violence
7,911⇌321 August
801⇌119 Lady
18,804⇌442 nazis
16,317⇌446 Heather
electoral 446⇌15,272
47,553⇌610 torches
JabberDuck 993⇌840,082
hatemonger 756⇌120,186
102,414⇌743 Antifa
opport 672⇌62,603
56,244⇌657 neo-nazis
Harambe 658⇌56,005
Michelle 261⇌3,115
50.2%—49.8%



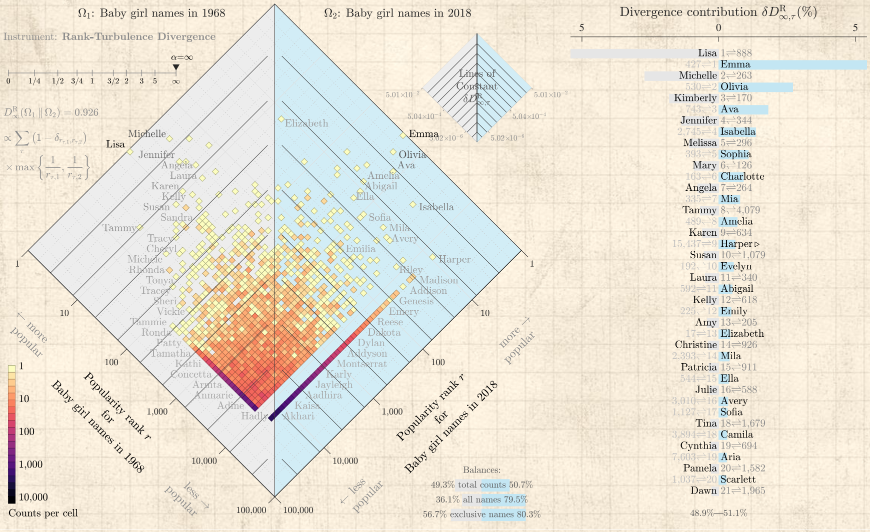
Ω1: Twitter on 2016/11/09
Ω2: Twitter on 2017/08/13
Instrument: Probability-Turbulence Divergence
α=1/12
0 1/4 1/2 3/4 1 3/2 2 3 5 ∞
$D^{\mathrm{P}}_{1/12}(\Omega_1 \,\|\, \Omega_2) = \sum_{\tau} \delta D^{\mathrm{P}}_{1/12,\tau}$
$= 13 \sum_{\tau} \left| p_{\tau,2}^{1/12} - p_{\tau,2}^{1/12} \right|^{12/13}$
Lines of Constant $\delta D^{\mathrm{P}}_{1/12,\tau}$
9.13×10^0
6.40×10^0
4.49×10^0
3.15×10^0
more talked about
less talked about
Probability p for Twitter on 2016/11/09
Probability p for Twitter on 2017/08/13
Counts per cell
1
10
100
1,000
10,000
100,000
1,000,000
Divergence contribution $\delta D^{\mathrm{P}}_{1/12,\tau}$ (×10^−4%)
1,552,865⇌672 Cvjetanovic ▷
1,552,865⇌1,116 FirstStarMagicAllStar ▷
1,552,865⇌1,474 KISSMARCREDy ▷
1,552,865⇌1,526 ForAllStarGames ▷
1,552,865⇌1,985 Kafeel ▷
1,552,865⇌2,021 Starbz ▷
◁ Bobama 2,423⇌1,537,471
◁ Oarack 2,425⇌1,537,471
◁ Un-Leashed 2,703⇌1,537,471
1,552,865⇌3,088 GSHDJHS ▷
1,552,865⇌3,099 Bodak ▷
◁ KiligTripSaBaguio 3,142⇌1,537,471
◁ Somali-American 3,229⇌1,537,471
◁ DICKASS 3,321⇌1,537,471
◁ Michellle 3,412⇌1,537,471
1,552,865⇌3,675 Eastwatch ▷
◁ Un-leashed 3,645⇌1,537,471
1,552,865⇌3,985 Heyer's ▷
◁ SCHOO 3,921⇌1,537,471
1,552,865⇌4,382 uncharacteristical ▷
1,552,865⇌4,518 callejones ▷
◁ misogy 4,328⇌1,537,471
1,552,865⇌4,723 TLC's ▷
1,552,865⇌4,913 SORIBADA ▷
◁ tRyNNa 4,660⇌1,537,471
◁ aLmoSt 4,671⇌1,537,471
1,552,865⇌5,246 tcas ▷
◁ Ruline 5,097⇌1,537,471
◁ Steininger 5,118⇌1,537,471
1,552,865⇌5,436 low-rise ▷
◁ climate-denying 5,191⇌1,537,471
1,552,865⇌5,667 CLITORIS ▷
1,552,865⇌5,682 Adityanath ▷
◁ lambo's 5,383⇌1,537,471
1,552,865⇌5,757 DeliHasret ▷
1,552,865⇌5,757 FikBel ▷
1,552,865⇌5,808 Walker-Peters ▷
◁ KBAT 5,617⇌1,537,471
1,552,865⇌6,046 UNIDAS ▷
◁ stunnered 5,653⇌1,537,471
49.9%—50.1%



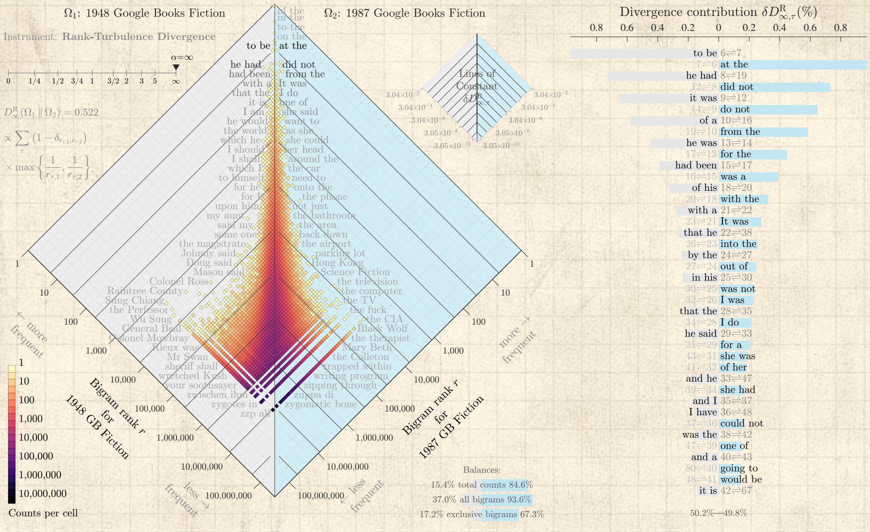
Ω1: Barro Colorado Island, 1985 Census
Ω2: Barro Colorado Island, 2015 Census
Instrument: Rank-Turbulence Divergence
α=0
0 1/4 1/2 3/4 1 3/2 2 3 5 ∞
$D^{R}_{0}(\Omega_1 \,\|\, \Omega_2) = \sum_\tau \delta D^{R}_{0,\tau}$
$= \sum_\tau \left| \ln \frac{r_{\tau,1}}{r_{\tau,2}} \right|$
= 0.077
Pombalia prunifolia
Faramea occidentalis
Trichilia tube...ata
Desmopsis panamensis
Oenocarpus mapora
Alseis blackiana
Mouriri myrtilloides
Swartzia simplex
Garcinia recondita
Psychotria hor...lis
Protium stevensonii
Sorocea affinis
Protium panamense
Tachigali panamensis
Coussarea curv...mia
Poulsenia armata
Eugenia galalonensis
Piper cordulatum
Calophyllum lo...ium
Guarea bullata
Inga marginata
Hasseltia floribunda
Palicourea gui...sis
Bactris major
Inga acuminata
Trophis caucana
Chamguava schippii
Piper cabagranum
Inga thibaudiana
Erythrina cost...sis
Cecropia obtusifolia
Bactris barronis
Cespedesia spa...ata
Bactris coloradonis
Trema integerrima
Cyathea petiolata
Cecropia longipes
Lines of Constant $\delta D^{R}_{0,\tau}$
6.91 4.61 2.3 0 0 2.3 4.61 6.91
1 10 100 1,000
← more abundant
less abundant →
more → abundant
← less abundant
Rank r for Barro Colorado Island, 1985 Census
Rank r for Barro Colorado Island, 2015 Census
1 10 100
Counts per cell
Balances:
51.5% total counts 48.5%
95.6% all types 92.5%
7.8% exclusive types 4.7%
Divergence contribution $\delta D^{R}_{0,\tau}$ (%)
4 3 2 1 0 1 2 3 4
Piper cordulatum 9⇌138
Poulsenia armata 14⇌53
65⇌22 Calophyllum longifolium
Psychotria horizontalis 8⇌23
46⇌16 Eugenia galalonensis
93⇌33 Palicourea guianensis
121⇌45 Inga acuminata
83⇌35 Cecropia insignis
39⇌17 Cupania seemannii
54⇌25 Xylopia macrantha
Hasseltia floribunda 37⇌77
Guarea bullata 34⇌70
Bactris barronis 137⇌269
127⇌65 Chamguava schippii
180⇌94 Inga thibaudiana
185⇌100 Cecropia obtusifolia
Ocotea whitei 44⇌81
Virola sebifera 22⇌40
78⇌43 Anaxagorea panamensis
Bactris major 48⇌86
16⇌9 Protium stevensonii
Tachigali panamensis 17⇌30
Piper cabagranum 98⇌170
Erythrina costaricensis 103⇌178
Piper culebranum 123⇌213
Guatteria lucens 29⇌50
31⇌18 Coussarea curvigemmia
Xylosma oligandra 97⇌165
Piper playablancanum 140⇌236
◁Bactris coloradonis 185⇌308
250⇌151 Cespedesia spathulata
Bactris coloniata 116⇌188
313⇌195 Trema integerrima ▷
Pouteria reticulata 30⇌48
204⇌128 Pourouma bicolor
Conostegia cinnamomea 85⇌135
19⇌12 Protium panamense
89⇌57 Chrysophyllum argenteum
20⇌13 Protium tenuifolium
74⇌49 Psychotria marginata
50.0%—50.0%



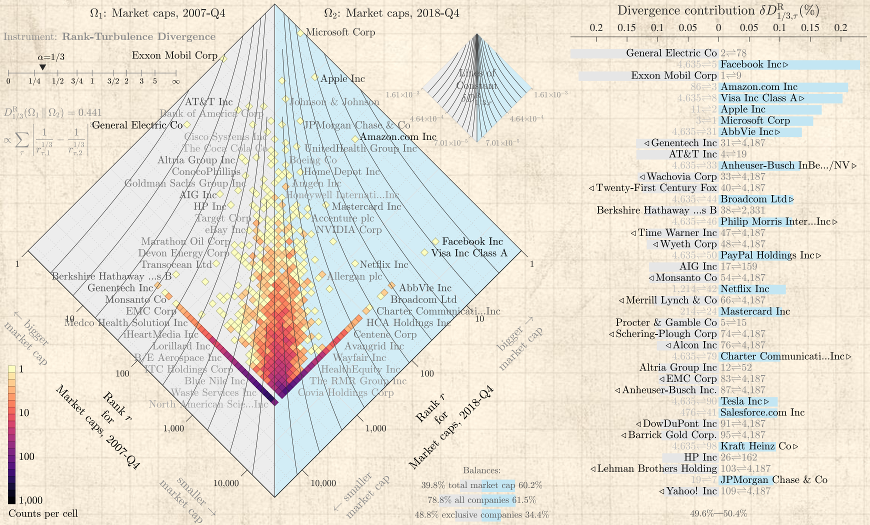
Ω1: Baby girl names in 1968
Ω2: Baby girl names in 2018
Instrument: Rank-Turbulence Divergence
α=∞
0 1/4 1/2 3/4 1 3/2 2 3 5 ∞
$D^{R}_{\infty}(\Omega_1 \| \Omega_2) = 0.926$
$\propto \sum_{\tau} (1 - \delta_{r_{\tau,1}, r_{\tau,2}})$
$\times \max\left\{\frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}}\right\}$
Lines of Constant $\delta D^{R}_{\infty,\tau}$
5.01×10^−2
5.04×10^−4
5.02×10^−6
Michelle
Lisa
Jennifer
Angela
Laura
Karen
Kelly
Susan
Sandra
Tammy
Tracy
Cheryl
Michele
Rhonda
Tonya
Tracey
Sheri
Vickie
Tammie
Ronda
Patty
Tamatha
Kathi
Concetta
Arnita
Anmarie
Adine
Hadly
Elizabeth
Emma
Olivia
Ava
Amelia
Abigail
Ella
Isabella
Sofia
Mila
Avery
Emilia
Harper
Riley
Madison
Addison
Genesis
Emery
Reese
Dakota
Dylan
Addyson
Montserrat
Karly
Jayleigh
Aadhira
Kaisa
Akhari
1
10
100
1,000
10,000
100,000
← more popular
more → popular
less → popular
← less popular
Popularity rank r for Baby girl names in 1968
Popularity rank r for Baby girl names in 2018
Counts per cell
1
10
100
1,000
10,000
Balances:
49.3% total counts 50.7%
36.1% all names 79.5%
56.7% exclusive names 80.3%
Divergence contribution $\delta D^{R}_{\infty,\tau}$ (%)
5 0 5
Lisa 1⇌888
427⇌1 Emma
Michelle 2⇌263
530⇌2 Olivia
Kimberly 3⇌170
1,743⇌3 Ava
Jennifer 4⇌344
2,745⇌4 Isabella
Melissa 5⇌296
393⇌5 Sophia
Mary 6⇌126
163⇌6 Charlotte
Angela 7⇌264
335⇌7 Mia
Tammy 8⇌4,079
489⇌8 Amelia
Karen 9⇌634
15,437⇌9 Harper ▷
Susan 10⇌1,079
192⇌10 Evelyn
Laura 11⇌340
592⇌11 Abigail
Kelly 12⇌618
225⇌12 Emily
Amy 13⇌205
17⇌13 Elizabeth
Christine 14⇌926
2,393⇌14 Mila
Patricia 15⇌911
544⇌15 Ella
Julie 16⇌588
3,010⇌16 Avery
1,127⇌17 Sofia
Tina 18⇌1,679
3,894⇌18 Camila
Cynthia 19⇌694
7,603⇌19 Aria
Pamela 20⇌1,582
1,037⇌20 Scarlett
Dawn 21⇌1,965
48.9%—51.1%



Ω₁: Baby girl names in 1968
Ω₂: Baby girl names in 2018
Instrument: Rank-Turbulence Divergence
α=∞
0 1/4 1/2 3/4 1 3/2 2 3 5 ∞
$D^{\mathrm{R}}_{\infty}(\Omega_1 \,\|\, \Omega_2) = 0.926$
$\propto \sum_{\tau} \left(1 - \delta_{r_{\tau,1}, r_{\tau,2}}\right)$
$\times \max\left\{\frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}}\right\}$
Lines of Constant $\delta D^{\mathrm{R}}_{\infty,\tau}$
5.01×10⁻² 5.04×10⁻⁴ 5.02×10⁻⁶
Michelle Lisa Jennifer Angela Laura Karen Kelly Susan Sandra Tammy Tracy Cheryl Michele Rhonda Tonya Tracey Sheri Vickie Tammie Ronda Patty Tamatha Kathi Concetta Arnita Anmarie Adine Hadly
Elizabeth Emma Olivia Ava Amelia Abigail Ella Isabella Sofia Mila Avery Emilia Harper Riley Madison Addison Genesis Emery Reese Dakota Dylan Addyson Montserrat Karly Jayleigh Aadhira Kaisa Akhari
1 10 100 1,000 10,000 100,000
← more popular
more → popular
less → popular
← less popular
Popularity rank r for Baby girl names in 1968
Popularity rank r for Baby girl names in 2018
1 10 100 1,000 10,000
Counts per cell
Balances:
49.3% total counts 50.7%
36.1% all names 79.5%
56.7% exclusive names 80.3%
Divergence contribution $\delta D^{\mathrm{R}}_{\infty,\tau}$ (%)
5 0 5
Lisa 1⇌888
427⇌1 Emma
Michelle 2⇌263
530⇌2 Olivia
Kimberly 3⇌170
1,743⇌3 Ava
Jennifer 4⇌344
2,745⇌4 Isabella
Melissa 5⇌296
393⇌5 Sophia
Mary 6⇌126
163⇌6 Charlotte
Angela 7⇌264
335⇌7 Mia
Tammy 8⇌4,079
489⇌8 Amelia
Karen 9⇌634
15,437⇌9 Harper ▷
Susan 10⇌1,079
192⇌10 Evelyn
Laura 11⇌340
592⇌11 Abigail
Kelly 12⇌618
225⇌12 Emily
Amy 13⇌205
17⇌13 Elizabeth
Christine 14⇌926
2,393⇌14 Mila
Patricia 15⇌911
544⇌15 Ella
Julie 16⇌588
3,010⇌16 Avery
1,127⇌17 Sofia
Tina 18⇌1,679
3,894⇌18 Camila
Cynthia 19⇌694
7,603⇌19 Aria
Pamela 20⇌1,582
1,037⇌20 Scarlett
Dawn 21⇌1,965
48.9%—51.1%

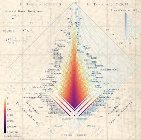


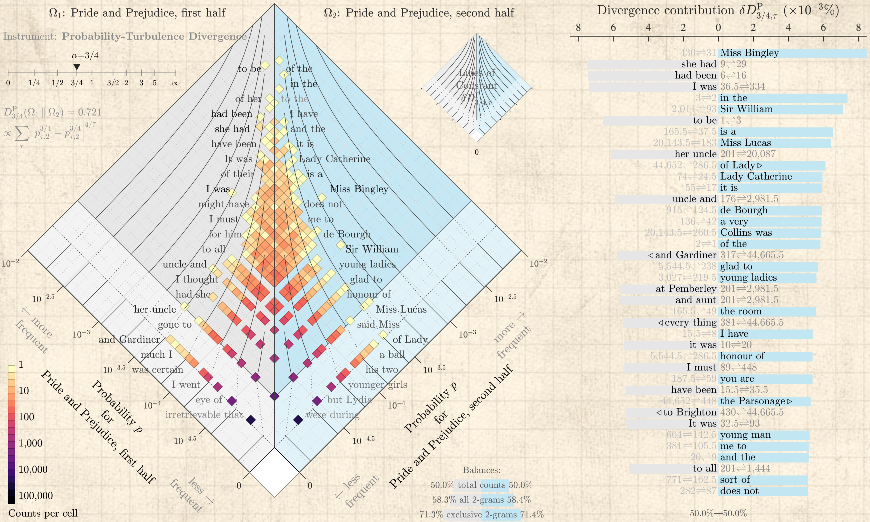
Ω1: 1948 Google Books Fiction
Ω2: 1987 Google Books Fiction
Instrument: Rank-Turbulence Divergence
α=∞
0 1/4 1/2 3/4 1 3/2 2 3 5 ∞
$D^{R}_{\infty}(\Omega_1 \,\|\, \Omega_2) = 0.522$
$\propto \sum_{\tau} \left(1-\delta_{r_{\tau,1},r_{\tau,2}}\right)$
$\times \max\left\{\frac{1}{r_{\tau,1}}, \frac{1}{r_{\tau,2}}\right\}$
Lines of Constant $\delta D^{R}_{\infty,\tau}$
Bigram rank r for 1948 GB Fiction
Bigram rank r for 1987 GB Fiction
← more frequent
more frequent →
less frequent →
← less frequent
Counts per cell
1
10
100
1,000
10,000
100,000
1,000,000
10,000,000
Balances:
15.4% total counts 84.6%
37.0% all bigrams 93.6%
17.2% exclusive bigrams 67.3%
Divergence contribution $\delta D^{R}_{\infty,\tau}$ (%)
to be 6⇌7
7⇌6 at the
he had 8⇌19
12⇌8 did not
it was 9⇌12
14⇌9 do not
of a 10⇌16
19⇌10 from the
he was 13⇌14
17⇌13 for the
had been 15⇌17
16⇌15 was a
of his 18⇌20
20⇌18 with the
with a 21⇌22
23⇌21 It was
that he 22⇌38
26⇌23 into the
by the 24⇌27
27⇌24 out of
in his 25⇌30
30⇌25 was not
32⇌26 I was
that the 28⇌35
34⇌28 I do
he said 29⇌33
31⇌29 for a
43⇌31 she was
41⇌32 of her
and he 33⇌47
39⇌34 she had
and I 35⇌37
I have 36⇌48
37⇌36 could not
was the 38⇌42
47⇌39 one of
and a 40⇌43
80⇌40 going to
48⇌41 would be
it is 42⇌67
50.2%—49.8%



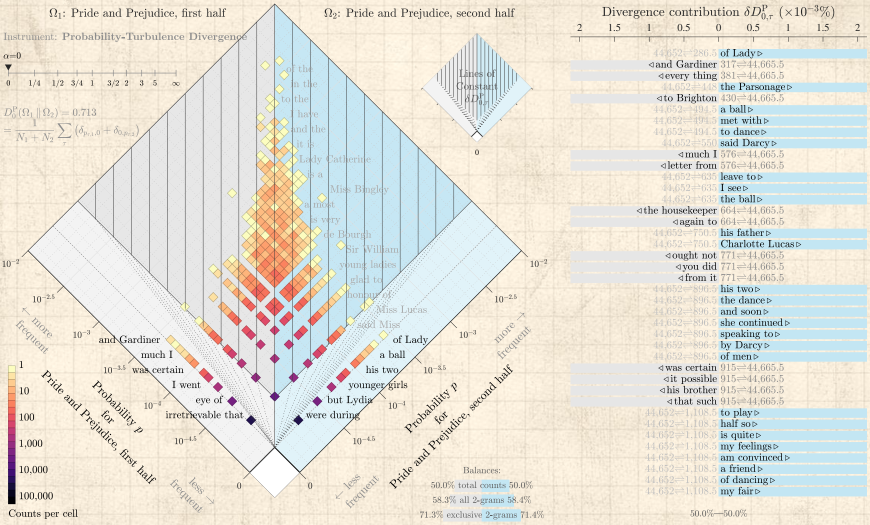
Ω_1: Market caps, 2007-Q4
Ω_2: Market caps, 2018-Q4
Instrument: **Rank-Turbulence Divergence**
α=1/3
0 1/4 1/2 3/4 1 3/2 2 3 5 ∞
$D^{R}_{1/3}(\Omega_1 \| \Omega_2) = 0.441$
$\propto \sum_{\tau} \left| \frac{1}{r_{\tau,1}^{1/3}} - \frac{1}{r_{\tau,2}^{1/3}} \right|$
Lines of Constant $\delta D^{R}_{1/3,\tau}$
1.61×10⁻³
4.64×10⁻⁴
7.01×10⁻⁵
Microsoft Corp
Exxon Mobil Corp
Apple Inc
AT&T Inc
Bank of America Corp
Johnson & Johnson
General Electric Co
JPMorgan Chase & Co
Amazon.com Inc
Cisco Systems Inc
The Coca Cola Co
UnitedHealth Group Inc
Altria Group Inc
Boeing Co
ConocoPhillips
Home Depot Inc
Goldman Sachs Group Inc
Amgen Inc
AIG Inc
Honeywell Internati...Inc
HP Inc
Mastercard Inc
Target Corp
Accenture plc
eBay Inc
NVIDIA Corp
Marathon Oil Corp
Facebook Inc
Devon Energy Corp
Visa Inc Class A
Transocean Ltd
Netflix Inc
Berkshire Hathaway ...s B
Allergan plc
Genentech Inc
AbbVie Inc
Monsanto Co
Broadcom Ltd
EMC Corp
Charter Communicati...Inc
Medco Health Solution Inc
HCA Holdings Inc
iHeartMedia Inc
Centene Corp
Lorillard Inc
Avangrid Inc
B/E Aerospace Inc
Wayfair Inc
ITC Holdings Corp
HealthEquity Inc
Blue Nile Inc
The RMR Group Inc
Waste Services Inc
Covia Holdings Corp
North American Scie...Inc
1
10
100
1,000
10,000
← bigger market cap
bigger market cap →
smaller market cap →
← smaller market cap
Rank r for Market caps, 2007-Q4
Rank r for Market caps, 2018-Q4
1
10
100
1,000
Counts per cell
Balances:
39.8% total market cap 60.2%
78.8% all companies 61.5%
48.8% exclusive companies 34.4%
Divergence contribution $\delta D^{R}_{1/3,\tau}$(%)
0.2 0.15 0.1 0.05 0 0.05 0.1 0.15 0.2
General Electric Co 2⇌78
4,635⇌5 Facebook Inc ▷
Exxon Mobil Corp 1⇌9
86⇌3 Amazon.com Inc
4,635⇌8 Visa Inc Class A ▷
11⇌2 Apple Inc
3⇌1 Microsoft Corp
4,635⇌31 AbbVie Inc ▷
◁ Genentech Inc 31⇌4,187
AT&T Inc 4⇌19
4,635⇌33 Anheuser-Busch InBe.../NV ▷
◁ Wachovia Corp 33⇌4,187
◁ Twenty-First Century Fox 40⇌4,187
4,635⇌44 Broadcom Ltd ▷
Berkshire Hathaway ...s B 38⇌2,331
4,635⇌46 Philip Morris Inter...Inc ▷
◁ Time Warner Inc 47⇌4,187
◁ Wyeth Corp 48⇌4,187
4,635⇌50 PayPal Holdings Inc ▷
AIG Inc 17⇌159
◁ Monsanto Co 54⇌4,187
1,214⇌42 Netflix Inc
◁ Merrill Lynch & Co 66⇌4,187
214⇌24 Mastercard Inc
Procter & Gamble Co 5⇌15
◁ Schering-Plough Corp 74⇌4,187
◁ Alcon Inc 76⇌4,187
4,635⇌79 Charter Communicati...Inc ▷
Altria Group Inc 12⇌52
◁ EMC Corp 83⇌4,187
◁ Anheuser-Busch Inc. 87⇌4,187
4,635⇌90 Tesla Inc ▷
476⇌41 Salesforce.com Inc
◁ DowDuPont Inc 91⇌4,187
◁ Barrick Gold Corp. 95⇌4,187
4,635⇌98 Kraft Heinz Co ▷
HP Inc 26⇌162
◁ Lehman Brothers Holding 103⇌4,187
19⇌7 JPMorgan Chase & Co
◁ Yahoo! Inc 109⇌4,187
49.6%—50.4%

Effect of subsampling:

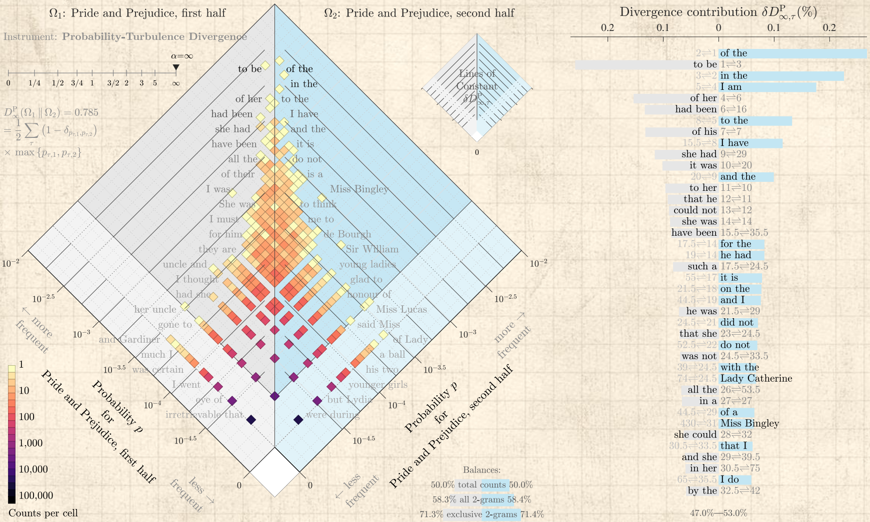




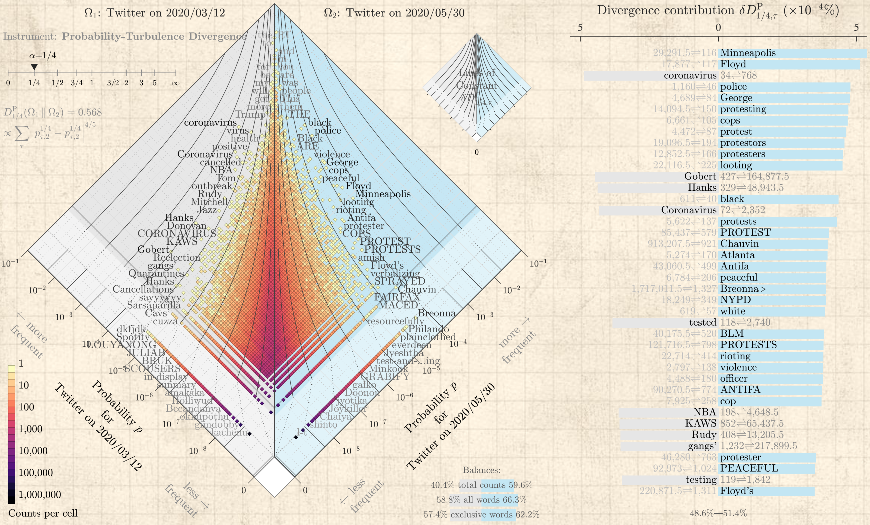
Ω1: Pride and Prejudice, first half
Ω2: Pride and Prejudice, second half
Instrument: Probability-Turbulence Divergence
α=3/4
0 1/4 1/2 3/4 1 3/2 2 3 5 ∞
$D^{P}_{3/4}(\Omega_1 \| \Omega_2) = 0.721$
$\propto \sum_{\tau} \left| p^{3/4}_{\tau,2} - p^{3/4}_{\tau,2} \right|^{4/7}$
Lines of Constant $\delta D^{P}_{3/4,\tau}$
to be
of the
in the
of her
to the
had been
I have
she had
and the
have been
it is
It was
Lady Catherine
of their
is a
I was
Miss Bingley
might have
does not
I must
me to
for him
de Bourgh
to all
Sir William
uncle and
young ladies
I thought
glad to
had she
honour of
her uncle
Miss Lucas
gone to
said Miss
and Gardiner
of Lady
much I
a ball
was certain
his two
I went
younger girls
eye of
but Lydia
irretrievable that
were during
10^{-2}
$10^{-2.5}$
10^{-3}
$10^{-3.5}$
10^{-4}
$10^{-4.5}$
0
← more frequent
more → frequent
less → frequent
← less frequent
Probability p for Pride and Prejudice, first half
Probability p for Pride and Prejudice, second half
1
10
100
1,000
10,000
100,000
Counts per cell
Balances:
50.0% total counts 50.0%
58.3% all 2-grams 58.4%
71.3% exclusive 2-grams 71.4%
Divergence contribution $\delta D^{P}_{3/4,\tau}$ ($\times 10^{-3}$%)
8 6 4 2 0 2 4 6 8
430⇌31 Miss Bingley
she had 9⇌29
had been 6⇌16
I was 36.5⇌334
3⇌2 in the
2,011⇌93 Sir William
to be 1⇌3
165.5⇌37.5 is a
20,143.5⇌183 Miss Lucas
her uncle 201⇌20,087
44,652⇌286.5 of Lady ▷
74⇌24.5 Lady Catherine
55⇌17 it is
uncle and 176⇌2,981.5
915⇌124.5 de Bourgh
136⇌42 a very
20,143.5⇌260.5 Collins was
2⇌1 of the
◁and Gardiner 317⇌44,665.5
5,544.5⇌238 glad to
3,027⇌219.5 young ladies
at Pemberley 201⇌2,981.5
and aunt 201⇌2,981.5
165.5⇌49 the room
◁every thing 381⇌44,665.5
15.5⇌8 I have
it was 10⇌20
5,544.5⇌286.5 honour of
I must 89⇌448
187.5⇌59 you are
have been 15.5⇌35.5
44,652⇌448 the Parsonage ▷
◁to Brighton 430⇌44,665.5
It was 32.5⇌93
664⇌142.5 young man
381⇌105.5 me to
20⇌9 and the
to all 201⇌1,444
771⇌162.5 sort of
282⇌87 does not
50.0%—50.0%



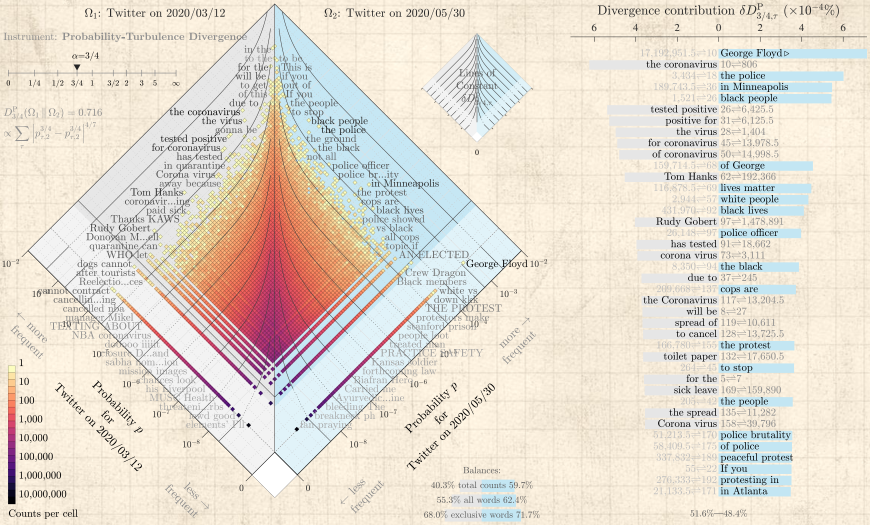
Ω1: Pride and Prejudice, first half
Ω2: Pride and Prejudice, second half
Instrument: Probability-Turbulence Divergence
α=0
0 1/4 1/2 3/4 1 3/2 2 3 5 ∞
$D_0^{\mathrm{P}}(\Omega_1 \| \Omega_2) = 0.713$
$= \frac{1}{N_1 + N_2} \sum_\tau \left(\delta_{p_{\tau,1},0} + \delta_{0,p_{\tau,2}} \right)$
Lines of Constant $\delta D^{\mathrm{P}}_{0,\tau}$
of the
in the
to the
I have
and the
it is
Lady Catherine
is a
Miss Bingley
a most
is very
de Bourgh
Sir William
young ladies
glad to
honour of
Miss Lucas
said Miss
and Gardiner
much I
was certain
I went
eye of
irretrievable that
of Lady
a ball
his two
younger girls
but Lydia
were during
← more frequent
more → frequent
less → frequent
← less frequent
Probability p for Pride and Prejudice, first half
Probability p for Pride and Prejudice, second half
10^{-2} $10^{-2.5}$ 10^{-3} $10^{-3.5}$ 10^{-4} $10^{-4.5}$ 0
1
10
100
1,000
10,000
100,000
Counts per cell
Balances:
50.0% total counts 50.0%
58.3% all 2-grams 58.4%
71.3% exclusive 2-grams 71.4%
Divergence contribution $\delta D^{\mathrm{P}}_{0,\tau}$ ($\times 10^{-3}$%)
2 1.5 1 0.5 0 0.5 1 1.5 2
44,652⇌286.5 of Lady ▹
◃ and Gardiner 317⇌44,665.5
◃ every thing 381⇌44,665.5
44,652⇌448 the Parsonage ▹
◃ to Brighton 430⇌44,665.5
44,652⇌494.5 a ball ▹
44,652⇌494.5 met with ▹
44,652⇌494.5 to dance ▹
44,652⇌550 said Darcy ▹
◃ much I 576⇌44,665.5
◃ letter from 576⇌44,665.5
44,652⇌635 leave to ▹
44,652⇌635 I see ▹
44,652⇌635 the ball ▹
◃ the housekeeper 664⇌44,665.5
◃ again to 664⇌44,665.5
44,652⇌750.5 his father ▹
44,652⇌750.5 Charlotte Lucas ▹
◃ ought not 771⇌44,665.5
◃ you did 771⇌44,665.5
◃ from it 771⇌44,665.5
44,652⇌896.5 his two ▹
44,652⇌896.5 the dance ▹
44,652⇌896.5 and soon ▹
44,652⇌896.5 she continued ▹
44,652⇌896.5 speaking to ▹
44,652⇌896.5 by Darcy ▹
44,652⇌896.5 of men ▹
◃ was certain 915⇌44,665.5
◃ it possible 915⇌44,665.5
◃ his brother 915⇌44,665.5
◃ that such 915⇌44,665.5
44,652⇌1,108.5 to play ▹
44,652⇌1,108.5 half so ▹
44,652⇌1,108.5 is quite ▹
44,652⇌1,108.5 my feelings ▹
44,652⇌1,108.5 am convinced ▹
44,652⇌1,108.5 a friend ▹
44,652⇌1,108.5 of dancing ▹
44,652⇌1,108.5 my fair ▹
50.0%—50.0%



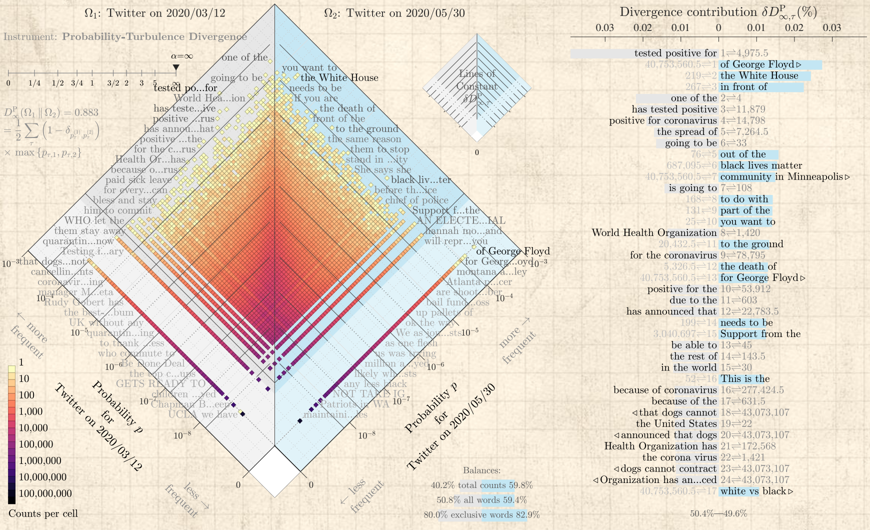
Ω1: Pride and Prejudice, first half
Ω2: Pride and Prejudice, second half
Instrument: Probability-Turbulence Divergence
α=∞
0 1/4 1/2 3/4 1 3/2 2 3 5 ∞
$D^{\mathrm{P}}_{\infty}(\Omega_1 \,\|\, \Omega_2) = 0.785$
$= \frac{1}{2} \sum_{\tau} \left(1 - \delta_{p_{\tau,1}, p_{\tau,2}}\right)$
$\times \max\{p_{\tau,1}, p_{\tau,2}\}$
Lines of Constant $\delta D^{\mathrm{P}}_{\infty,\tau}$
0
to be
of her
had been
she had
have been
all the
of their
I was
She was
I must
for him
they are
uncle and
I thought
had she
her uncle
gone to
and Gardiner
much I
was certain
I went
eye of
irretrievable that
of the
in the
to the
I have
and the
it is
do not
is a
Miss Bingley
to think
me to
de Bourgh
Sir William
young ladies
glad to
honour of
Miss Lucas
said Miss
of Lady
a ball
his two
younger girls
but Lydia
were during
10^{-2}
$10^{-2.5}$
10^{-3}
$10^{-3.5}$
10^{-4}
$10^{-4.5}$
0
← more frequent
less → frequent
more → frequent
← less frequent
Probability p for Pride and Prejudice, first half
Probability p for Pride and Prejudice, second half
1
10
100
1,000
10,000
100,000
Counts per cell
Balances:
50.0% total counts 50.0%
58.3% all 2-grams 58.4%
71.3% exclusive 2-grams 71.4%
Divergence contribution $\delta D^{\mathrm{P}}_{\infty,\tau}(\%)$
0.2 0.1 0 0.1 0.2
2⇌1 of the
to be 1⇌3
3⇌2 in the
5⇌4 I am
of her 4⇌6
had been 6⇌16
8⇌5 to the
of his 7⇌7
15.5⇌8 I have
she had 9⇌29
it was 10⇌20
20⇌9 and the
to her 11⇌10
that he 12⇌11
could not 13⇌12
she was 14⇌14
have been 15.5⇌35.5
17.5⇌14 for the
19⇌14 he had
such a 17.5⇌24.5
55⇌17 it is
21.5⇌18 on the
44.5⇌19 and I
he was 21.5⇌29
24.5⇌21 did not
that she 23⇌24.5
52.5⇌22 do not
was not 24.5⇌33.5
39⇌24.5 with the
74⇌24.5 Lady Catherine
all the 26⇌53.5
in a 27⇌27
44.5⇌29 of a
430⇌31 Miss Bingley
she could 28⇌32
30.5⇌33.5 that I
and she 29⇌39.5
in her 30.5⇌75
65⇌35.5 I do
by the 32.5⇌42
47.0%—53.0%



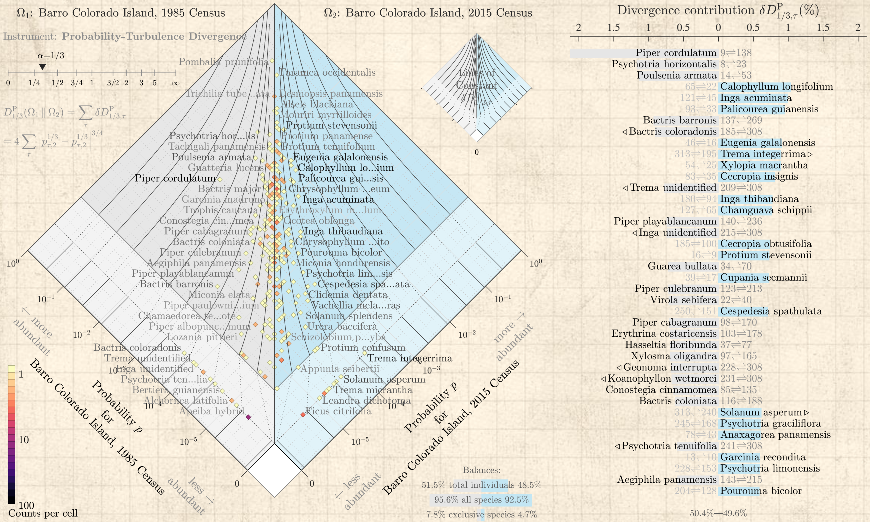
Ω1: Twitter on 2020/03/12
Ω2: Twitter on 2020/05/30
Instrument: Probability-Turbulence Divergence
α=1/4
$D^{P}_{1/4}(\Omega_1 \| \Omega_2) = 0.568$
$\propto \sum_\tau \left| p_{\tau,1}^{1/4} - p_{\tau,2}^{1/4} \right|^{4/5}$
Lines of Constant $\delta D^{P}_{1/4,\tau}$
Probability p for Twitter on 2020/03/12
Probability p for Twitter on 2020/05/30
← more frequent
more → frequent
less → frequent
← less frequent
Counts per cell
Balances:
40.4% total counts 59.6%
58.8% all words 66.3%
57.4% exclusive words 62.2%
Divergence contribution $\delta D^{P}_{1/4,\tau}$ (×10^{-4}%)
29,291.5⇌116 Minneapolis
17,877⇌117 Floyd
coronavirus 34⇌768
1,160⇌46 police
4,689⇌84 George
14,094.5⇌150 protesting
6,661⇌105 cops
4,472⇌87 protest
19,096.5⇌194 protestors
12,852.5⇌166 protesters
22,116.5⇌225 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 ▹
18,249⇌349 NYPD
619⇌57 white
tested 118⇌2,740
40,175.5⇌520 BLM
121,716.5⇌798 PROTESTS
22,714⇌414 rioting
2,797⇌138 violence
4,488⇌180 officer
90,270.5⇌774 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⇌763 protester
92,973⇌1,024 PEACEFUL
testing 119⇌1,842
220,871.5⇌1,311 Floyd's
48.6%—51.4%



Ω1: Twitter on 2020/03/12
Ω2: Twitter on 2020/05/30
Instrument: Probability-Turbulence Divergence
α=3/4
0 1/4 1/2 3/4 1 3/2 2 3 5 ∞
$D^{P}_{3/4}(\Omega_1 \| \Omega_2) = 0.716$
$\propto \sum_{\tau} \left| p_{\tau,2}^{3/4} - p_{\tau,2}^{3/4} \right|^{4/7}$
Lines of Constant $\delta D^{P}_{3/4,\tau}$
Divergence contribution $\delta D^{P}_{3/4,\tau}$ (×10⁻⁴%)
6 4 2 0 2 4 6
17,192,951.5⇌10 George Floyd ▹
the coronavirus 10⇌806
3,434⇌18 the police
189,743.5⇌36 in Minneapolis
1,521⇌26 black people
tested positive 26⇌6,425.5
positive for 31⇌6,125.5
the virus 28⇌1,404
for coronavirus 45⇌13,978.5
of coronavirus 50⇌14,998.5
159,714.5⇌68 of George
Tom Hanks 62⇌192,366
116,878.5⇌69 lives matter
2,944⇌57 white people
431,970⇌92 black lives
Rudy Gobert 97⇌1,478,891
26,148⇌97 police officer
has tested 91⇌18,662
corona virus 73⇌3,111
8,350⇌94 the black
due to 37⇌245
209,668⇌137 cops are
the Coronavirus 117⇌13,204.5
will be 8⇌27
spread of 119⇌10,611
to cancel 128⇌13,725.5
166,780⇌155 the protest
toilet paper 132⇌17,650.5
264⇌45 to stop
for the 5⇌7
sick leave 169⇌159,890
205⇌42 the people
the spread 135⇌11,282
Corona virus 158⇌39,796
51,213.5⇌170 police brutality
58,409.5⇌175 of police
337,832⇌189 peaceful protest
55⇌22 If you
276,333⇌192 protesting in
21,133.5⇌171 in Atlanta
51.6%—48.4%
Balances:
40.3% total counts 59.7%
55.3% all words 62.4%
68.0% exclusive words 71.7%
Counts per cell
1 10 100 1,000 10,000 100,000 1,000,000 10,000,000
Probability p for Twitter on 2020/03/12
Probability p for Twitter on 2020/05/30
← more frequent
more → frequent
less → frequent
← less frequent



Ω_1: Twitter on 2020/03/12
Ω_2: Twitter on 2020/05/30
Instrument: Probability-Turbulence Divergence
α=∞
0 1/4 1/2 3/4 1 3/2 2 3 5 ∞
$D^{P}_{\infty}(\Omega_1 \| \Omega_2) = 0.883$
$= \frac{1}{2}\sum_{\tau}\left(1 - \delta_{p^{(1)}_{\tau}, p^{(2)}_{\tau}}\right)$
$\times \max\{p_{\tau,1}, p_{\tau,2}\}$
Lines of Constant $\delta D^{P}_{\infty,\tau}$
Counts per cell
1
10
100
1,000
10,000
100,000
1,000,000
10,000,000
100,000,000
Probability p for Twitter on 2020/03/12
Probability p for Twitter on 2020/05/30
← more frequent
more frequent →
less frequent →
← less frequent
Balances:
40.2% total counts 59.8%
50.8% all words 59.4%
80.0% exclusive words 82.9%
Divergence contribution $\delta D^{P}_{\infty,\tau}$(%)
0.03 0.02 0.01 0 0.01 0.02 0.03
tested positive for 1⇌4,975.5
40,753,560.5⇌1 of George Floyd ▷
219⇌2 the White House
267⇌3 in front of
one of the 2⇌4
has tested positive 3⇌11,879
positive for coronavirus 4⇌14,798
the spread of 5⇌7,264.5
going to be 6⇌33
76⇌5 out of the
687,095⇌6 black lives matter
40,753,560.5⇌7 community in Minneapolis ▷
is going to 7⇌108
168⇌8 to do with
131⇌9 part of the
25⇌10 you want to
World Health Organization 8⇌1,420
20,432.5⇌11 to the ground
for the coronavirus 9⇌78,795
5,326.5⇌12 the death of
40,753,560.5⇌13 for George Floyd ▷
positive for the 10⇌53,912
due to the 11⇌603
has announced that 12⇌22,783.5
199⇌14 needs to be
3,040,697⇌15 Support from the
be able to 13⇌45
the rest of 14⇌143.5
in the world 15⇌30
52⇌16 This is the
because of coronavirus 16⇌277,424.5
because of the 17⇌631.5
◁ that dogs cannot 18⇌43,073,107
the United States 19⇌22
◁ announced that dogs 20⇌43,073,107
Health Organization has 21⇌172,568
the corona virus 22⇌1,421
◁ dogs cannot contract 23⇌43,073,107
◁ Organization has an...ced 24⇌43,073,107
40,753,560.5⇌17 white vs black ▷
50.4%—49.6%



Ω1: Barro Colorado Island, 1985 Census
Ω2: Barro Colorado Island, 2015 Census
Instrument: Probability-Turbulence Divergence
α=1/3
0 1/4 1/2 3/4 1 3/2 2 3 5 ∞
$D^{\mathrm{P}}_{1/3}(\Omega_1 \,\|\, \Omega_2) = \sum_\tau \delta D^{\mathrm{P}}_{1/3,\tau}$
$= 4 \sum_\tau \left| p_{\tau,2}^{1/3} - p_{\tau,2}^{1/3} \right|^{3/4}$
Lines of Constant $\delta D^{\mathrm{P}}_{1/3,\tau}$
Pombalia prunifolia
Faramea occidentalis
Trichilia tube...ata
Desmopsis panamensis
Alseis blackiana
Mouriri myrtilloides
Protium stevensonii
Psychotria hor...lis
Protium panamense
Tachigali panamensis
Protium tenuifolium
Poulsenia armata
Eugenia galalonensis
Guatteria lucens
Calophyllum lo...ium
Piper cordulatum
Palicourea gui...sis
Bactris major
Chrysophyllum ...eum
Garcinia madruno
Inga acuminata
Trophis caucana
Erythroxylum m...lum
Conostegia cin...mea
Ocotea oblonga
Piper cabagranum
Inga thibaudiana
Bactris coloniata
Chrysophyllum ...ito
Piper culebranum
Pourouma bicolor
Aegiphila panamensis
Miconia hondurensis
Piper playablancanum
Psychotria lim...sis
Bactris barronis
Cespedesia spa...ata
Miconia elata
Clidemia dentata
Piper paulowni...ium
Vachellia mela...ras
Chamaedorea te...ote
Solanum splendens
Piper alboponc...mum
Urera baccifera
Lozania pittieri
Schizolobium p...yba
Bactris coloradonis
Protium confusum
Trema unidentified
Trema integerrima
Inga unidentified
Appunia seibertii
Psychotria ten...lia
Solanum asperum
Bertiera guianensis
Trema micrantha
Alchornea latifolia
Leandra dichotoma
Apeiba hybrid
Ficus citrifolia
10^0 10^{-1} 10^{-2} 10^{-3} 10^{-4} 10^{-5} 0
← more abundant
less → abundant
more → abundant
← less abundant
Probability p for Barro Colorado Island, 1985 Census
Probability p for Barro Colorado Island, 2015 Census
1
10
100
Counts per cell
Balances:
51.5% total individuals 48.5%
95.6% all species 92.5%
7.8% exclusive species 4.7%
Divergence contribution $\delta D^{\mathrm{P}}_{1/3,\tau}$ (%)
2 1.5 1 0.5 0 0.5 1 1.5 2
Piper cordulatum 9⇌138
Psychotria horizontalis 8⇌23
Poulsenia armata 14⇌53
65⇌22 Calophyllum longifolium
121⇌45 Inga acuminata
93⇌33 Palicourea guianensis
Bactris barronis 137⇌269
◁Bactris coloradonis 185⇌308
46⇌16 Eugenia galalonensis
313⇌195 Trema integerrima▷
54⇌25 Xylopia macrantha
83⇌35 Cecropia insignis
◁Trema unidentified 209⇌308
180⇌94 Inga thibaudiana
127⇌65 Chamguava schippii
Piper playablancanum 140⇌236
◁Inga unidentified 215⇌308
185⇌100 Cecropia obtusifolia
16⇌9 Protium stevensonii
Guarea bullata 34⇌70
39⇌17 Cupania seemannii
Piper culebranum 123⇌213
Virola sebifera 22⇌40
250⇌151 Cespedesia spathulata
Piper cabagranum 98⇌170
Erythrina costaricensis 103⇌178
Hasseltia floribunda 37⇌77
Xylosma oligandra 97⇌165
◁Geonoma interrupta 228⇌308
◁Koanophyllon wetmorei 231⇌308
Conostegia cinnamomea 85⇌135
Bactris coloniata 116⇌188
313⇌240 Solanum asperum▷
245⇌168 Psychotria graciliflora
78⇌43 Anaxagorea panamensis
◁Psychotria tenuifolia 241⇌308
13⇌10 Garcinia recondita
228⇌153 Psychotria limonensis
Aegiphila panamensis 143⇌215
204⇌128 Pourouma bicolor
50.4%—49.6%

Flipbooks:

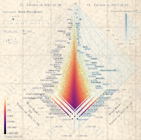
- Twitter:
  instrument-flipbook-1-rank-div.pdf
  instrument-flipbook-2-probability-div.pdf
  instrument-flipbook-3-gen-entropy-div.pdf
- Market caps:
  instrument-flipbook-4-marketcaps-6years-rank-div.pdf
- Baby names:
  instrument-flipbook-5-babynames-girls-50years-rank-div.pdf
  instrument-flipbook-6-babynames-boys-50years-rank-div.pdf
- Google books:
  instrument-flipbook-7-google-books-onegrams-rank-div.pdf
  instrument-flipbook-8-google-books-bigrams-rank-div.pdf
  instrument-flipbook-9-google-books-trigrams-rank-div.pdf

Flipbooks:

- Pride and Prejudice, 1-grams
- Pride and Prejudice, 2-grams
- Pride and Prejudice, 3-grams
- Twitter, 1-grams
- Twitter, 2-grams
- Twitter, 3-grams
- Barro Colorado Island

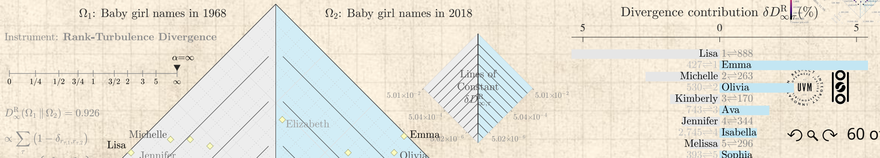
Code:

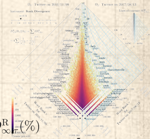
https://gitlab.com/compstorylab/allotaxonometer



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
- Maybe one day: Online tunable version of rank-turbulence divergence (plus many other instruments)



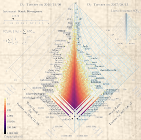


References I

[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

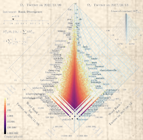
[3] M.-M. Deza and E. Deza.
Dictionary of Distances.
Elsevier, 2006.



References II

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

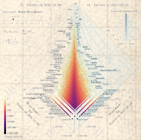
[5] P. S. Dodds, J. R. Minot, M. V. Arnold, T. Alshaabi, J. L. Adams, D. R. Dewhurst, T. J. Gray, M. R. Frank, A. J. Reagan, and C. M. Danforth.
Allotaxonometry and rank-turbulence divergence: A universal instrument for comparing complex systems, 2020.
Available online at https://arxiv.org/abs/2002.09770. pdf



References III

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

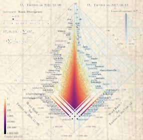


References IV

[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

[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

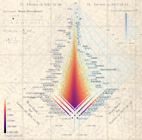


References V

[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

[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.

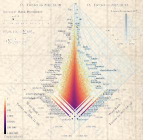


References VI

[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.

[16] 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.



References VII

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

[18] 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