Allotaxonometry

Last updated: 2023/05/22, 06:32:26 CEST

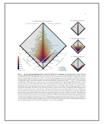
Principles of Complex Systems, Vols. 1, 2, & 3D
CSYS/MATH 300, 303, & 394, 2022–2023 | @pocsvox

Prof. Peter Sheridan Dodds | @peterdodds

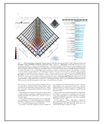
Computational Story Lab | Vermont Complex Systems Center
Santa Fe Institute | University of Vermont

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

Foundational papers:

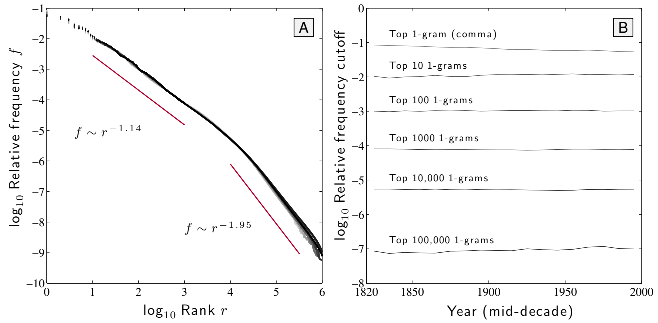


"Allotaxonometry and rank-turbulence divergence: A universal instrument for comparing complex systems"
Dodds et al.,
, 2020. [9]



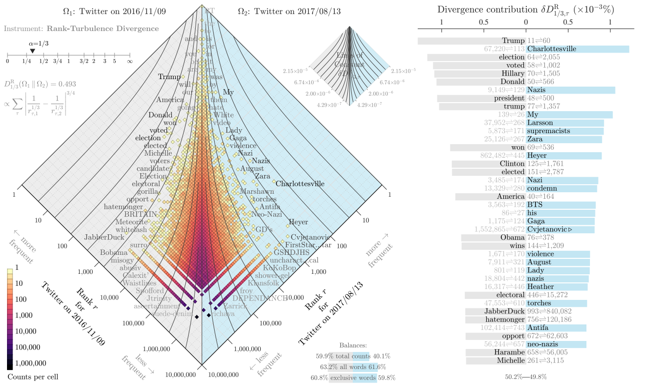
"Probability-turbulence divergence: A tunable allotaxonometric instrument for comparing heavy-tailed categorical distributions"
Dodds et al.,
, 2020. [11]

"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. [25]



Outline

Goal—Understand this:

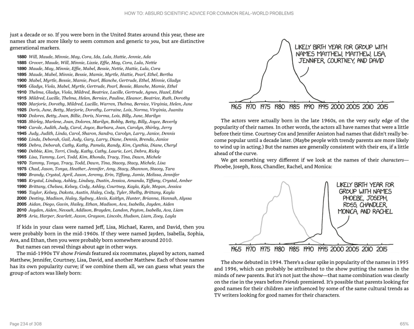


Basic science = Describe + Explain:

- 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 [23]

Baby names, much studied: [23]



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

For language, Zipf's law has two scaling regimes: [34]

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

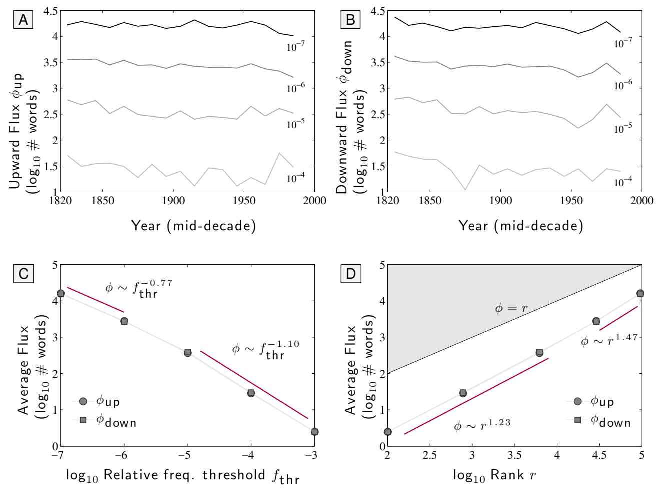
When comparing two texts, define Lexical turbulence as flux of words across a frequency threshold:

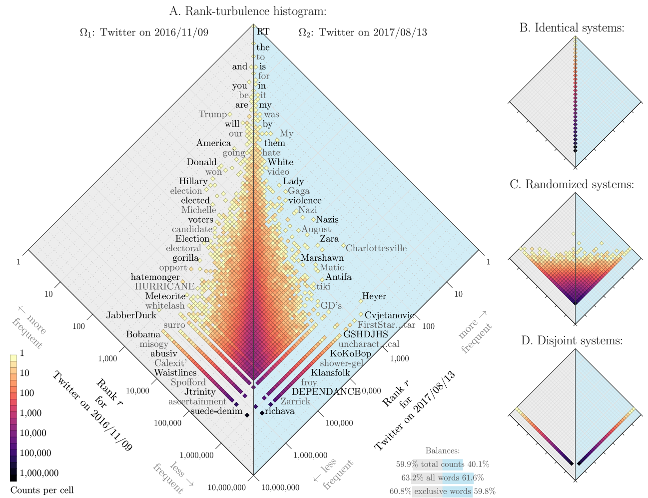
$$\phi \sim \begin{cases} f_{\text{thr}}^{-\mu} \text{ for } f_{\text{thr}} \ll f_b, \\ f_{\text{thr}}^{-\mu'} \text{ for } f_{\text{thr}} \gg f_b, \end{cases}$$

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 $\nu' \simeq 1.47$.

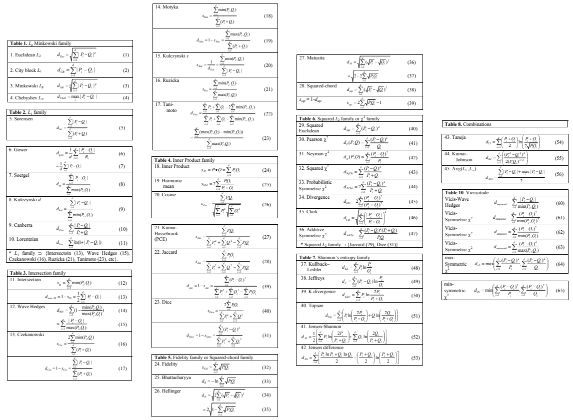


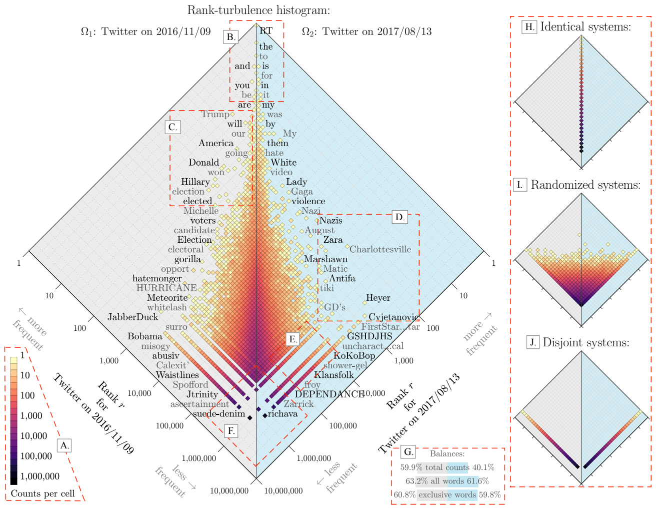


Exclusive types:

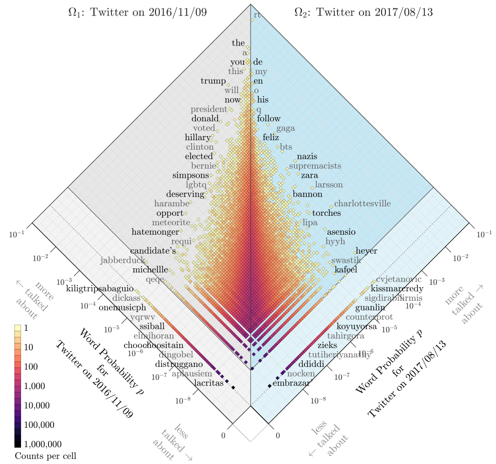
- We call types that are present in one system only 'exclusive types'.
- When warranted, we will use expressions of the form $\Omega^{(1)}$-exclusive and $\Omega^{(2)}$-exclusive to indicate to which system an exclusive type belongs.

Quite the festival:





Probability-turbulence histogram:



Shannon tried to slow things down in 1956:

"The bandwagon"
Claude E Shannon,
IRE Transactions on Information Theory, **2**, 3, 1956. [30]

- "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."

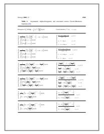
G. Balances:

59.9% total counts 40.1%

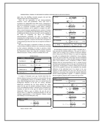
63.2% all words 61.6%

60.8% exclusive words 59.8%

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. [6]



"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. [3]

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

- 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 ⊃ {Intersection (13), Wave Hedges (15), Czekanowski (16), Ruzicka (21), Tanimoto (23), etc}.

- Information theoretic sortings are more opaque
- No tunability

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

Table 2. L_1 family		
5. Sørensen	$d_{sor} = \frac{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert}{\sum_{i=1}^{d} (P_i + Q_i)}$	(5)
6. Gower	$d_{gow} = \frac{1}{d} \sum_{i=1}^{d} \frac{\lvert P_i - Q_i \rvert}{R_i}$	(6)
	$= \frac{1}{d} \sum_{i=1}^{d} \lvert P_i - Q_i \rvert$	(7)
7. Soergel	$d_{sg} = \frac{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert}{\sum_{i=1}^{d} \max(P_i, Q_i)}$	(8)
8. Kulczynski d	$d_{kul} = \frac{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert}{\sum_{i=1}^{d} \min(P_i, Q_i)}$	(9)
9. Canberra	$d_{Can} = \sum_{i=1}^{d} \frac{\lvert P_i - Q_i \rvert}{P_i + Q_i}$	(10)
10. Lorentzian	$d_{Lor} = \sum_{i=1}^{d} \ln(1 + \lvert P_i - Q_i \rvert)$	(11)
* L_1 family $\supset$ {Intersectoin (13), Wave Hedges (15), Czekanowski (16), Ruzicka (21), Tanimoto (23), etc}.		

- Shannon's Entropy:

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

- Kullback-Liebler (KL) divergence:

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

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

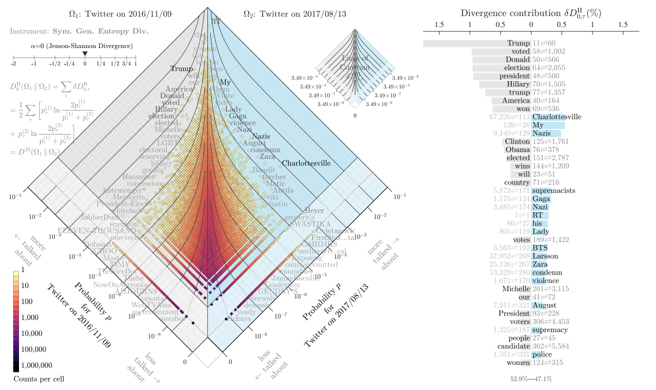
- Jensen-Shannon divergence (JSD): [19, 13, 24, 3]

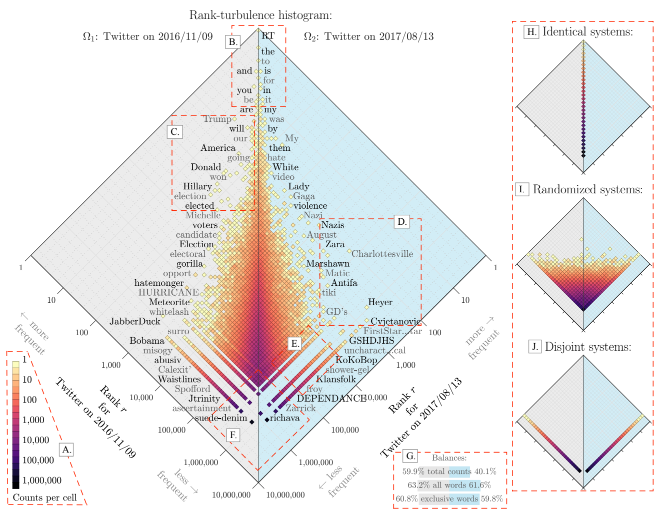
$$D^{\mathrm{JS}}(P_1 \,||\, P_2)$$
$$= \tfrac{1}{2} D^{\mathrm{KL}}\left(P_1 \,\middle\|\, \tfrac{1}{2}[P_1 + P_2]\right) + \tfrac{1}{2} D^{\mathrm{KL}}\left(P_2 \,\middle\|\, \tfrac{1}{2}[P_1 + P_2]\right)$$
$$= \frac{1}{2} \sum_{\tau \in R_{1,2;\alpha}} \left(p_{1,\tau} \log_2 \frac{p_{1,\tau}}{\frac{1}{2}[p_{1,\tau}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

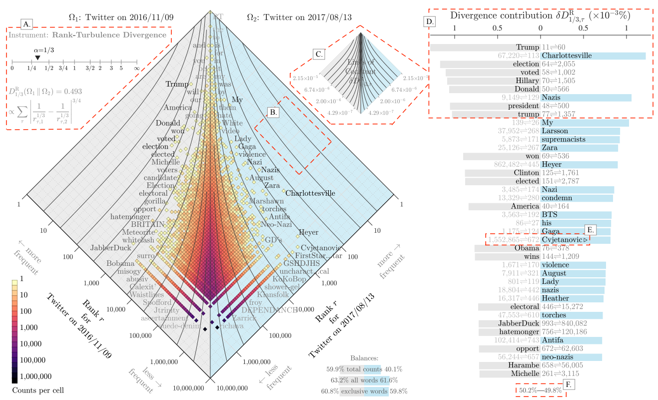
- Involving a third intermediate averaged system means JSD is now finite: $0 \le D^{\mathrm{JS}}(P_1 \,||\, P_2) \le 1$.
- Generalized entropy divergence: [6]

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

Produces JSD when $\alpha \to 0$.







Desirable rank-turbulence divergence features:

1. Rank-based.
2. Symmetric.
3. Semi-positive: $D^{\mathrm{R}}_{\alpha}(\Omega_1 \,||\, \Omega_2) \ge 0$.
4. Linearly separable, for interpretability.
5. Subsystem applicable: Ranked lists of any principled subset may be equally well compared (e.g., hashtags on Twitter, stock prices of a certain sector, etc.).
6. Turbulence-handling: Suited for systems with rank-ordered component size distribution that are heavy-tailed.
7. Scalable: Allow for sensible comparisons across system sizes.
8. Tunable.
9. Story-finding: Features 1–8 combine to show which component types are most 'important'

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 $\alpha \to 0$, 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

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

- The leading order term is:

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

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

- Oops.
- But the insides look nutritious:

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

is a nicely interpretable log-ratio of ranks.

Some reworking:

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

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

Rank-turbulence divergence:

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

Normalization:

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

Rank-turbulence divergence:

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

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

General normalization:

- Iif the Zipf distributions are disjoint, then in $\Omega^{(1)}$'s merged ranking, the rank of all $\Omega^{(2)}$ types will be $r = N_1 + \frac{1}{2}N_2$, where N_1 and N_2 are the number of distinct types in each system.
- Similarly, $\Omega^{(2)}$'s merged ranking will have all of $\Omega^{(1)}$'s types in last place with rank $r = N_2 + \frac{1}{2}N_1$.
- The normalization is then:

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

Limit of $\alpha \to 0$:

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

where

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

- Largest rank ratios dominate.

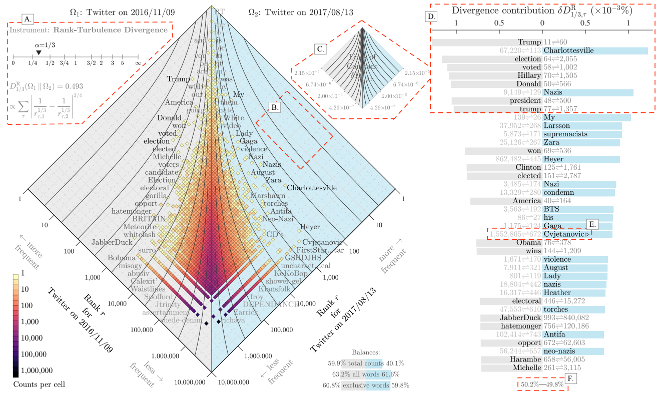
Limit of $\alpha \to \infty$:

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

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:

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

- 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:

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

A plenitude of distances
Rank-turbulence divergence
Probability-turbulence divergence
Explorations
Stories
Mechanics of Fame
Superspreading
Lexical Ultrafame
Turbulent times
References

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 $1/\alpha$.

A plenitude of distances
Rank-turbulence divergence
Probability-turbulence divergence
Explorations
Stories
Mechanics of Fame
Superspreading
Lexical Ultrafame
Turbulent times
References

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.

A plenitude of distances
Rank-turbulence divergence
Probability-turbulence divergence
Explorations
Stories
Mechanics of Fame
Superspreading
Lexical Ultrafame
Turbulent times
References

Combine these cases into a single expression:

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

- 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,

A plenitude of distances
Rank-turbulence divergence
Probability-turbulence divergence
Explorations
Stories
Mechanics of Fame
Superspreading
Lexical Ultrafame
Turbulent times
References

Type contribution ordering for the limit of α=0

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

A plenitude of distances
Rank-turbulence divergence
Probability-turbulence divergence
Explorations
Stories
Mechanics of Fame
Superspreading
Lexical Ultrafame
Turbulent times
References

Limit of α=∞ for probability-turbulence divergence

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

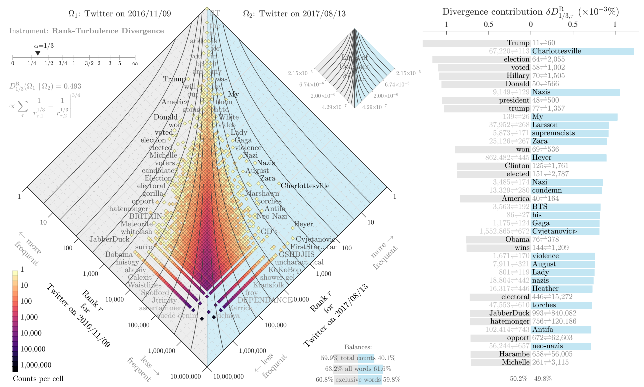
where

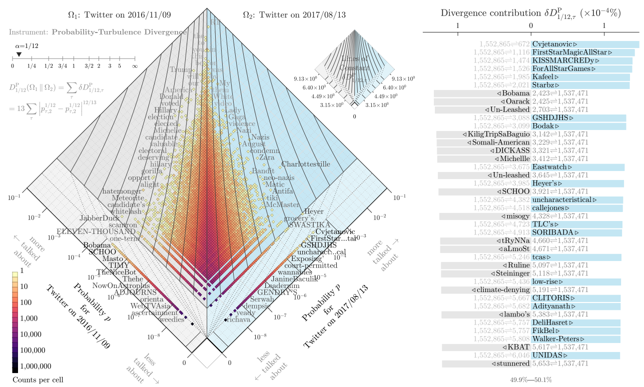
$$\mathcal{N}^{\mathsf{P}}_{1,2;\infty} = \sum_{\tau\in R_{1,2;\infty}}\left(p_{\tau,1}+p_{\tau,2}\right) = 1+1 = 2. \tag{22}$$

A plenitude of distances
Rank-turbulence divergence
Probability-turbulence divergence
Explorations
Stories
Mechanics of Fame
Superspreading
Lexical Ultrafame
Turbulent times
References

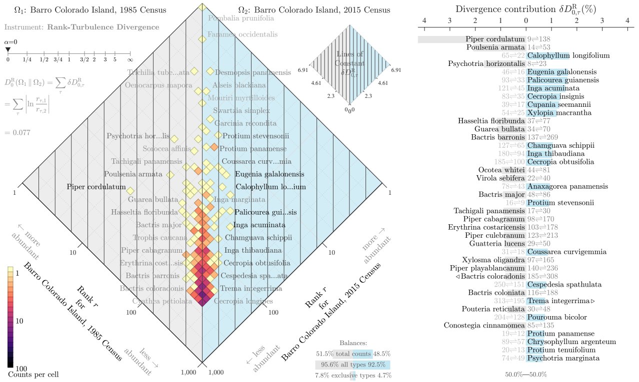
Connections for PTD:

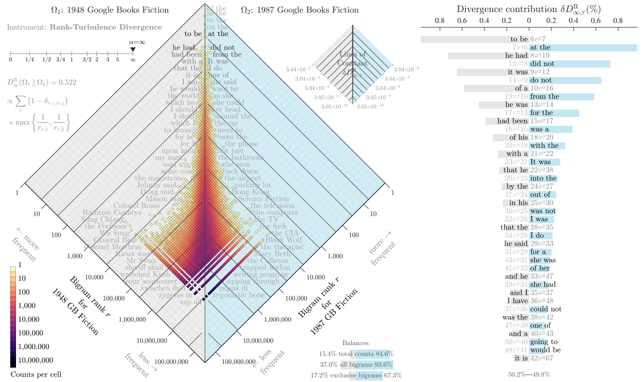
- $\alpha = 0$: Similarity measure Sørensen-Dice coefficient [8, 31, 20], F_1 score of a test's accuracy [32, 29].
- $\alpha = 1/2$: Hellinger distance [16] and Mautusita distance [21].
- $\alpha = 1$: Many including all $L^{(p)}$-norm type constructions.
- $\alpha = \infty$: Motyka distance [7].

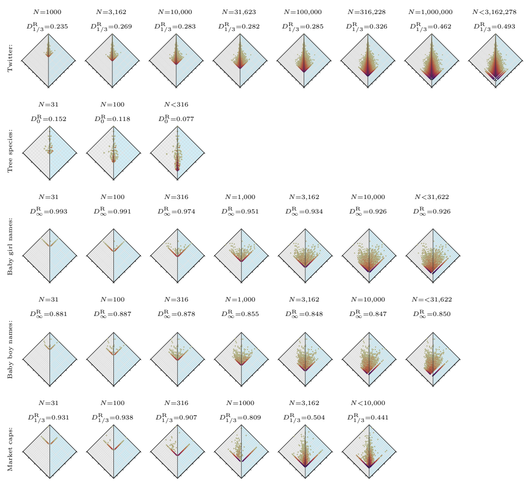


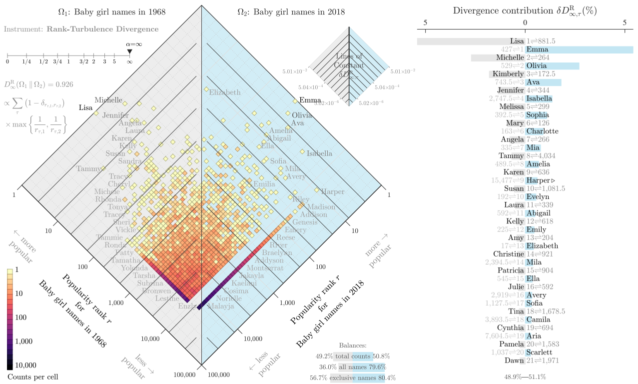


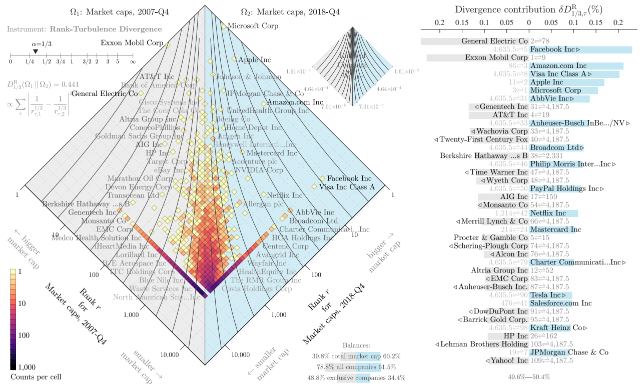
Effect of subsampling:

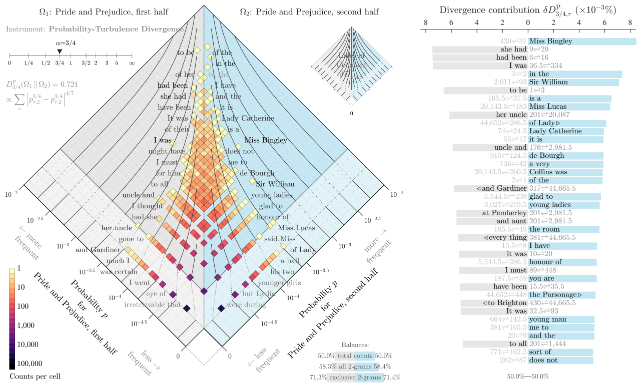


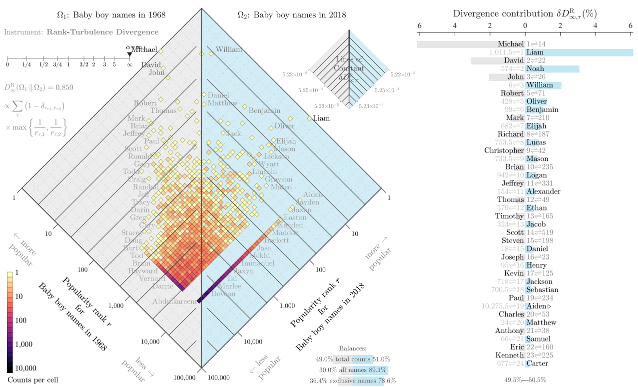












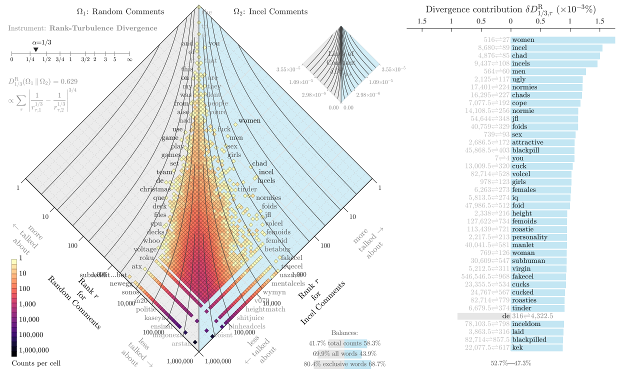
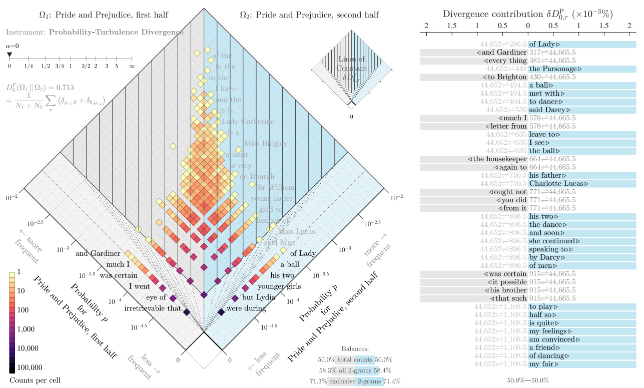
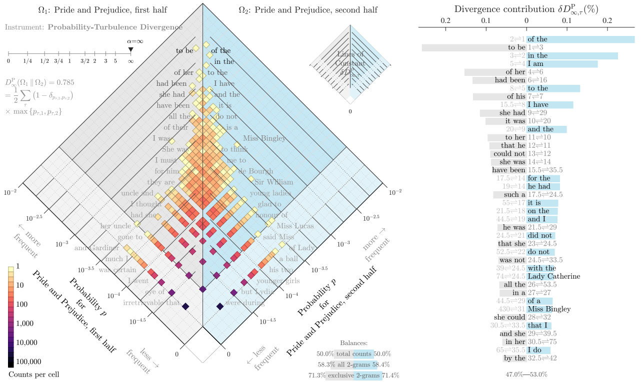
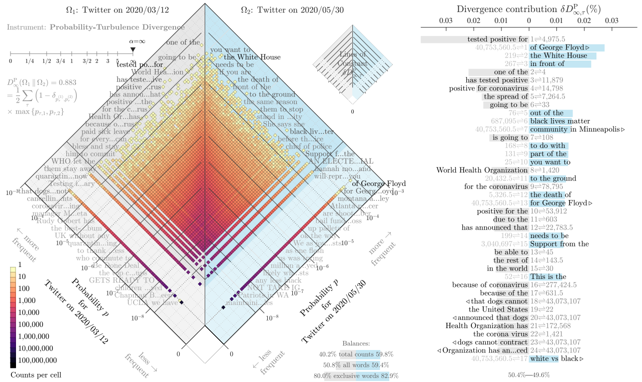


FIG. 8. **Rank-turbulence divergence allotaxonograph [34] of word rank distributions in the incel vs random comment corpora.** The rank-rank histogram on the left shows the density of words by their rank in the incel comments corpus against their rank in the random comments corpus. Words at the top of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

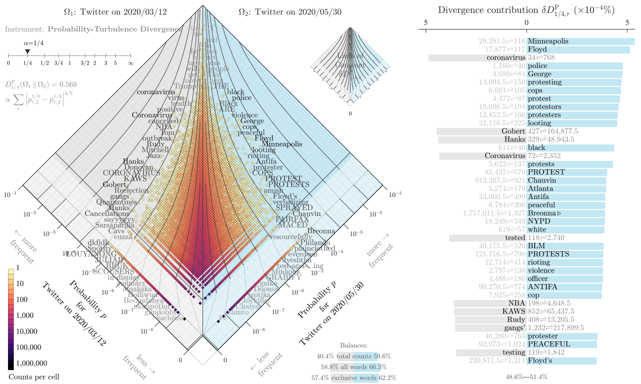


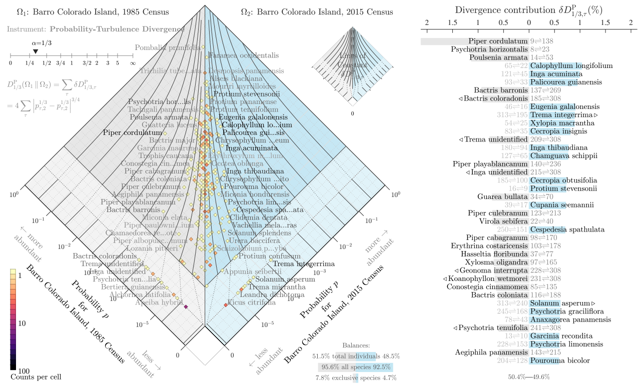




Flipbooks for PTD:

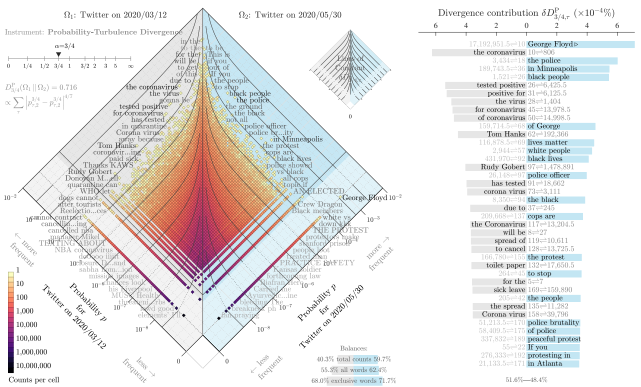
- Jane Austen:
  - Pride and Prejudice, 1-grams
  - Pride and Prejudice, 2-grams
  - Pride and Prejudice, 3-grams
- Social media:
  - Twitter, 1-grams
  - Twitter, 2-grams
  - Twitter, 3-grams
- Ecology:
  - Barro Colorado Island





Code:

https://gitlab.com/compstorylab/allotaxonometer

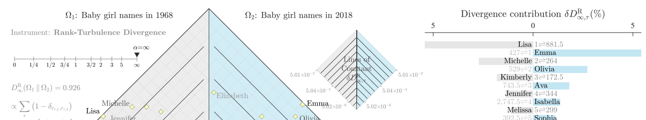


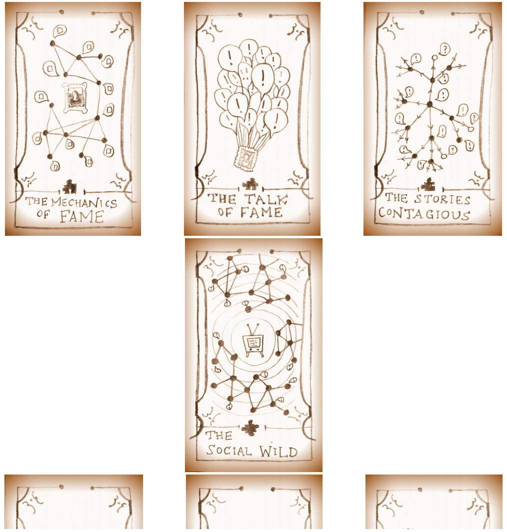
Flipbooks for RTD:

- 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

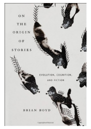
Claims, exaggerations, reminders:

- 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)

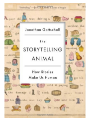




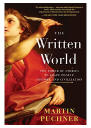
The everywhereness of algorithms and stories:



"On the Origin of Stories: Evolution, Cognition, and Fiction" by Brian Boyd (2010).[2]

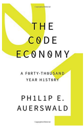


"The Storytelling Animal: How Stories Make Us Human" by Jonathan Gottschall (2013).[15]

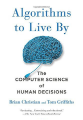


"The Written World: How Literature Shaped Civilization" by Martin Puchner (2017).[27]

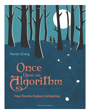
Algorithms, recipes, stories, ...



"The Code Economy: A Forty-Thousand Year History" by Philip E Auerswald (2017).[1]



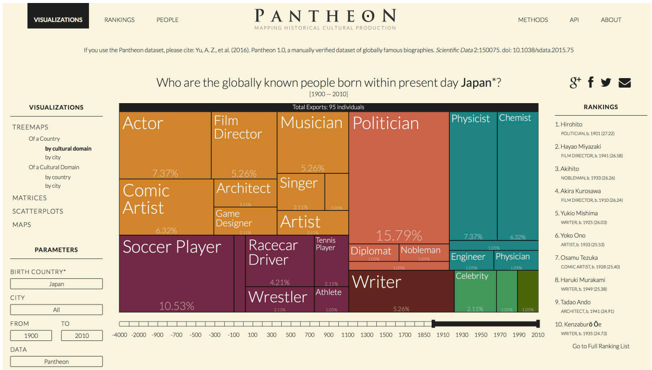
"Algorithms to Live By" by Christian and Griffiths (2016).[5]



"Once Upon an Algorithm" by Martin Erwig (2017).[14]

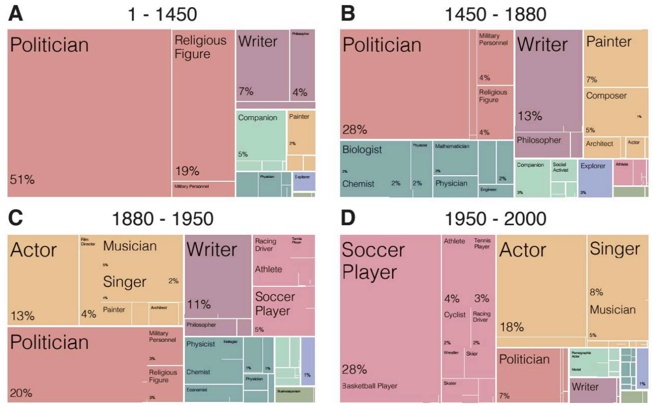
Also: Numerical Recipes in C[26] and How to Bake π[4]

The famous are storytellers—Japan:



For people born 1950–

http://pantheon.media.mit.edu/treemap/country_exports/JP/all/1900/2010/H15/pantheon



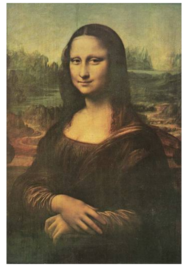
https://www.media.mit.edu/projects/pantheon-new/overview/

Super Survival of the Stories:

The Desirability of Storytellers, The Atlantic, Ed Yong, 2017-12-05.

- Study of Agta, Filipino hunter-gatherers.
- Storytelling valued well above all other skills including hunting.
- Stories encode prosocial norms such as cooperation.
- Like the best stories, the best storytellers reproduce more successfully.

The most famous painting in the world:



The dismal predictive powers of editors ......

Twelve ...

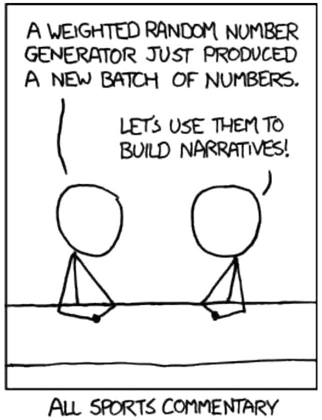
The completely unpredicted fall of Eastern Europe:

Timur Kuran:[18] "Now Out of Never: The Element of Surprise in the East European Revolution of 1989"

We understand bushfire stories:

1. Sparks start fires.
2. System properties control a fire's spread.
3. But for three reasons, we make two mistakes about Social Fires ...

Reason 1—We are Homo Narrativus.



http://xkcd.com/904/

Reason 2—"We are all individuals."

Archival footage:

- Individual narratives are not enough to understand distributed, networked minds.

Reason 3—We are spectacular imitators.

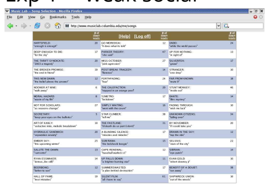
BBC/David Attenborough.

Mistake 1:
Success is due to intrinsic properties

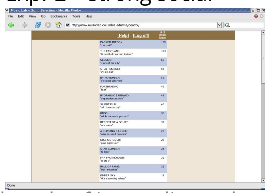
See "Becoming Mona Lisa" by David Sassoon

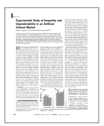
48 songs
30k participants

Exp 1— weak social



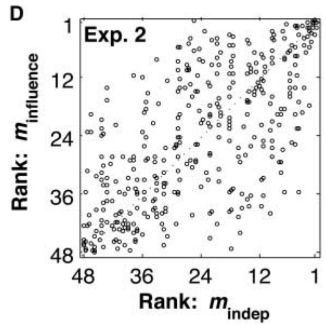
Exp. 2—strong social





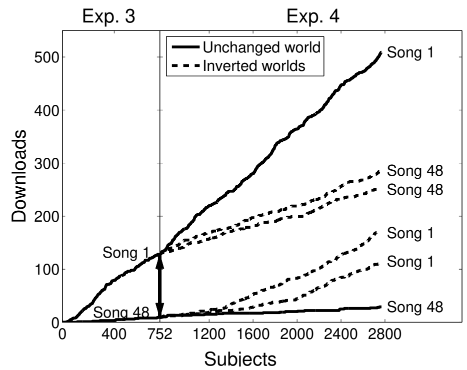
"An experimental study of inequality and unpredictability in an artificial cultural market"
Salganik, Dodds, and Watts, Science, **311**, 854–856, 2006. [28]

Resolving the paradox:



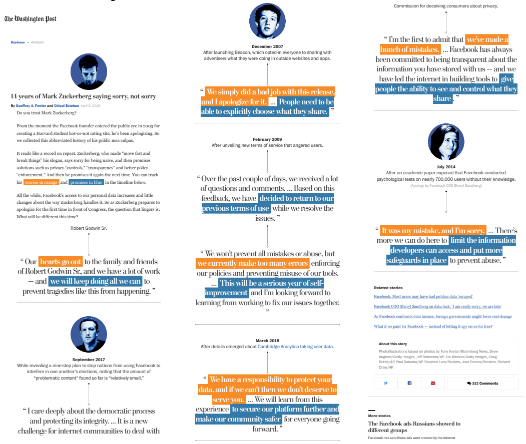
Increased social awareness leads to Stronger inequality + Less predictability.

Payola/Deceptive advertising hurts us all:



"Mistake" 2:
Seeing success is 'due to social' and wanting to say 'all your interactions are belong to us'

"This is truly the last time, believe me"


The Washington Post

Business • Analysis

14 years of Mark Zuckerberg saying sorry, not sorry

By Geoffrey A. Fowler and Chiqui Esteban April 9, 2018

Do you trust Mark Zuckerberg?

From the moment the Facebook founder entered the public eye in 2003 for creating a Harvard student hot-or-not rating site, he's been apologizing. So we collected this abbreviated history of his public mea culpas.

It reads like a record on repeat. Zuckerberg, who made "move fast and break things" his slogan, says sorry for being naive, and then promises solutions such as privacy "controls," "transparency" and better policy "enforcement." And then he promises it again the next time. You can track his sorries in orange and promises in blue in the timeline below.

All the while, Facebook's access to our personal data increases and little changes about the way Zuckerberg handles it. So as Zuckerberg prepares to apologize for the first time in front of Congress, the question that lingers is: What will be different this time?

Robert Godwin Sr.

" Our hearts go out to the family and friends of Robert Godwin Sr., and we have a lot of work — and we will keep doing all we can to prevent tragedies like this from happening. "

September 2017
While revealing a nine-step plan to stop nations from using Facebook to interfere in one another's elections, noting that the amount of "problematic content" found so far is "relatively small."

" I care deeply about the democratic process and protecting its integrity. ... It is a new challenge for internet communities to deal with

December 2007
After launching Beacon, which opted-in everyone to sharing with advertisers what they were doing in outside websites and apps.

" We simply did a bad job with this release, and I apologize for it. ... People need to be able to explicitly choose what they share. "

February 2009
After unveiling new terms of service that angered users.

" Over the past couple of days, we received a lot of questions and comments. ... Based on this feedback, we have decided to return to our previous terms of use while we resolve the issues. "

" We won't prevent all mistakes or abuse, but we currently make too many errors enforcing our policies and preventing misuse of our tools. ... This will be a serious year of self-improvement and I'm looking forward to learning from working to fix our issues together. "

March 2018
After details emerged about Cambridge Analytica taking user data.

" We have a responsibility to protect your data, and if we can't then we don't deserve to serve you. ... We will learn from this experience to secure our platform further and make our community safer for everyone going forward. "

Commission for deceiving consumers about privacy.

" I'm the first to admit that we've made a bunch of mistakes. ... Facebook has always been committed to being transparent about the information you have stored with us — and we have led the internet in building tools to give people the ability to see and control what they share. "

July 2014
After an academic paper exposed that Facebook conducted psychological tests on nearly 700,000 users without their knowledge.
(Apology by Facebook COO Sheryl Sandberg)

" It was my mistake, and I'm sorry. ... There's more we can do here to limit the information developers can access and put more safeguards in place to prevent abuse. "

Related stories
Facebook: Most users may have had publica data 'scraped'
Facebook COO Sheryl Sandberg on data leak: 'I am really sorry, we are late'
As Facebook confronts data misuse, foreign governments might force real change
What if we paid for Facebook — instead of letting it spy on us for free?

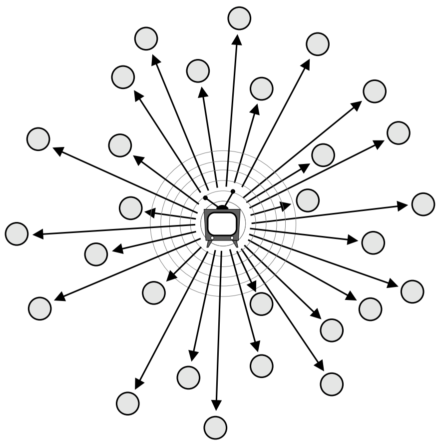
About this story
Photoillustrations based on photos by Tony Avelar/Bloomberg News, Drew Angerer/Getty Images, Jeff Roberson/AP, Jim Watson/Getty Images, Craig Ruttle/AP, Paul Sakuma/AP, Stephen Lam/Reuters, Jose Gomez/Reuters, Richard Drew/AP.

332 Comments

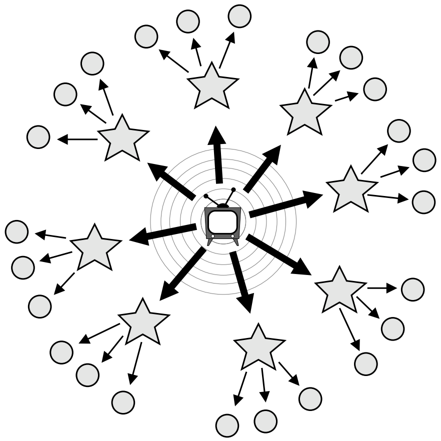
More stories
The Facebook ads Russians showed to different groups
Facebook has said these ads were created by the Internet

WaPo article

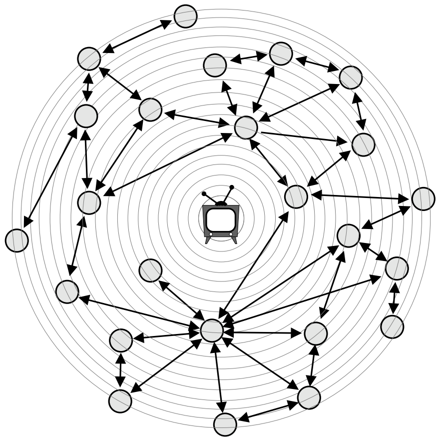
The hypodermic model of influence:



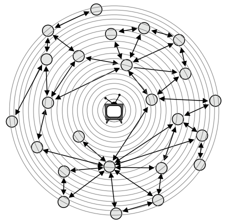
The two step model of influence:[17]



The network model of influence:

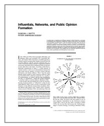


The network model of influence:



How superspreading works:

Many interconnected, average, trusting people must benefit from both receiving and sharing a message far from its source.

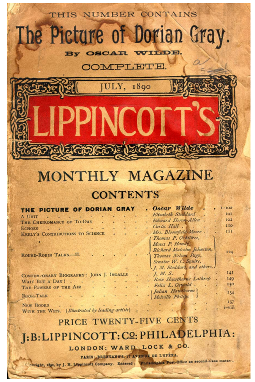


"Influentials, Networks, and Public Opinion Formation"
Watts and Dodds,
J. Consum. Res., **34**, 441–458, 2007.[33]

Etymological clarity:

- Fate—from the Latin *fatus*: meaning "spoken".
- Fate is talk that has been done. "It is written", fore-tell, pre-dict.
- "There is no such thing as fate, only the story of fate."
- Destiny is probablistic.
- Fame—from the Latin *fāma*: meaning "to talk."
- Fame is inherently the social discussion about the thing, not the thing itself.
- Renown: Repeatedly named, talked about. Old French renon, from re- + non ("name").
- Réclame. "Clamo"—Proto-Indo-European: "to shout" (again). Connected to "lowing".

Oscar Wilde, The Picture of Dorian Gray: Raw Fame



"There is only one thing in the world

worse than being talked about,

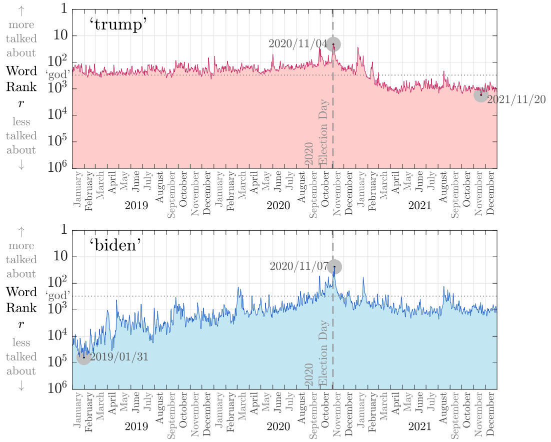
and that is

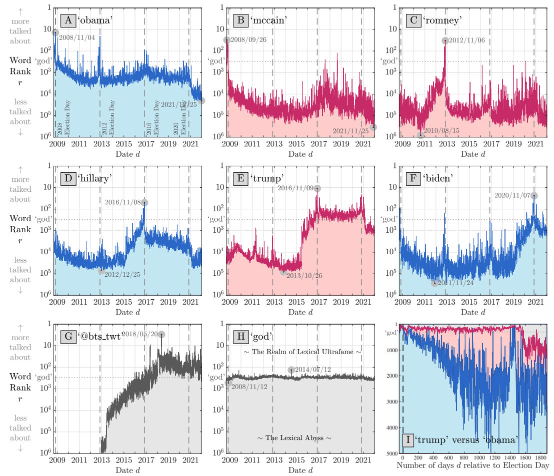
not being talked about."

"Fame and Ultrafame: Measuring and comparing daily levels of 'being talked about' for United States' presidents, their rivals, God, countries, and K-pop"
Dodds et al.,
Available online at https://arxiv.org/abs/1910.00149, 2019.[10]

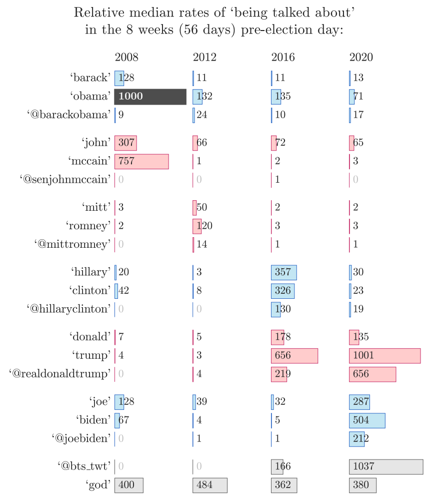
"Computational timeline reconstruction of the stories surrounding Trump: Story turbulence, narrative control, and collective chronopathy"
Dodds et al.,
, 2020.[12]

- POTUSometer with the Smorgasdashbord: http://compstorylab.org/potusometer/
- Stories surrounding Trump: http://compstorylab.org/trumpstoryturbulence/



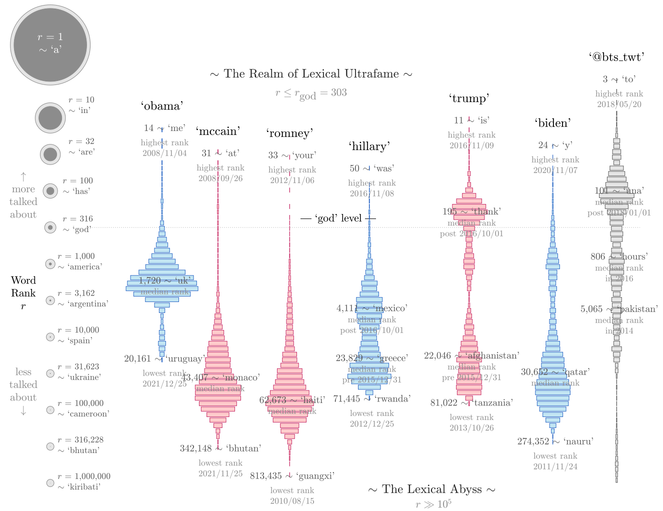


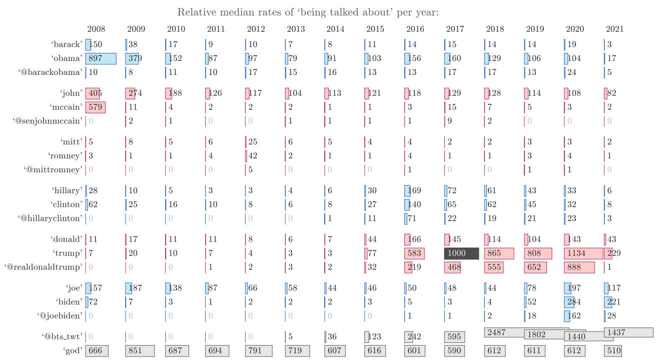
2011 Whitehouse Correspondents' Dinner



Ultraframe:
Nobody expects ~~the Spanish Inquisition~~ K-pop:

Vox (2019-04-17):
BTS, the band that changed K-pop, explained



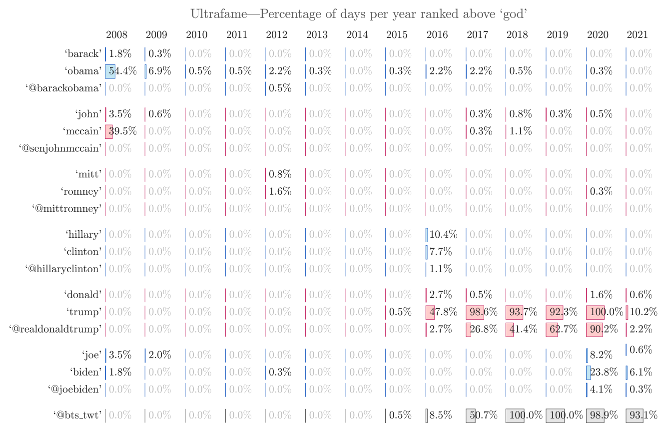


Telegnomics

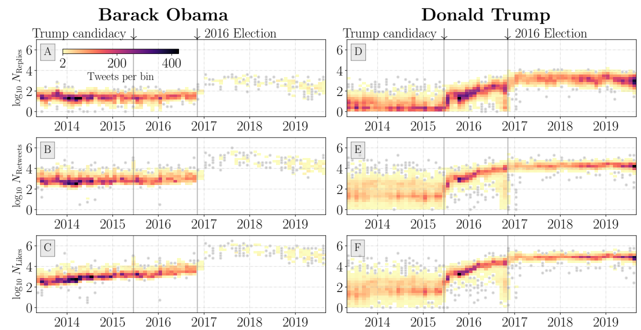
Distant reading by smashing texts into storyons:

```
cd ~/work/stories/2019-10story-turbulence-trump/
261G
more updateall.sh
file names:
compute_rank_turbulence_divergence_sweep_the_leg

Zip files:
zless 2018-01-06/1grams/en_*.tar.tsv
zless 2021-01-05/1grams/en_*.tar.tsv
zless 2021-01-06/1grams/en_*.tar.tsv
zless 2021-01-07/1grams/en_*.tar.tsv
```

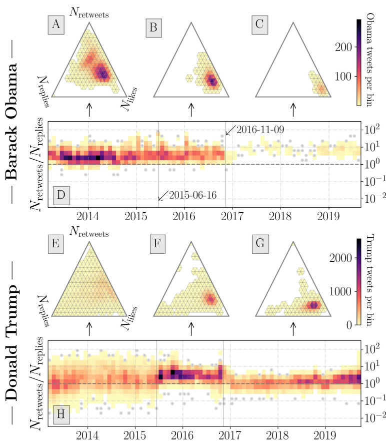


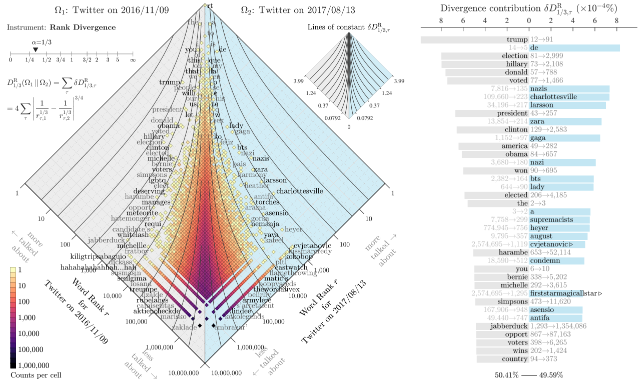
Ratiometrics:



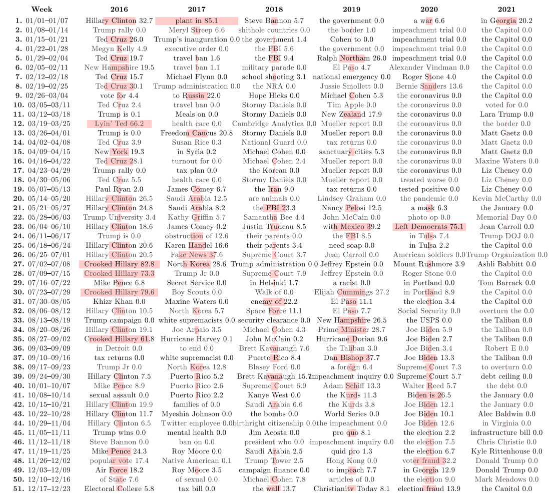
"Ratioing the President: An exploration of public engagement with Obama and Trump on Twitter," Minot et al., 2020 [22]

Ratiometrics:

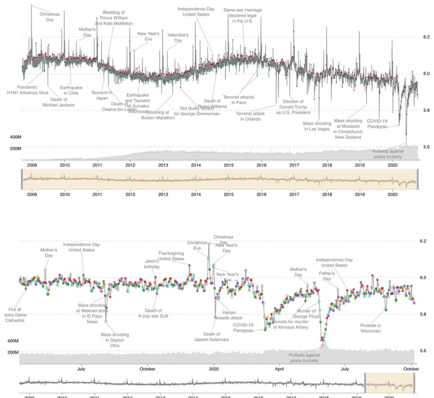




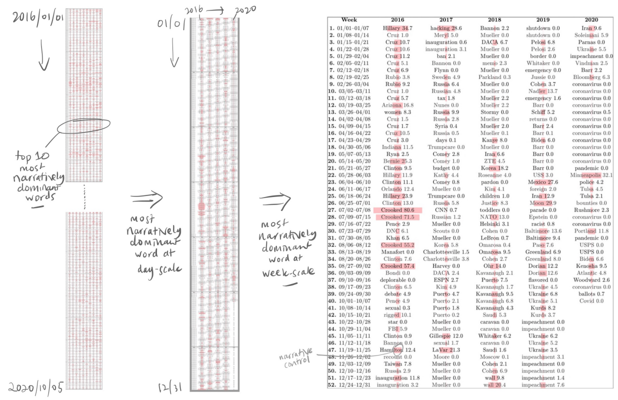
Allotaxonometry—
the comparison of complex systems:
http://compstorylab.org/allotaxonometry/



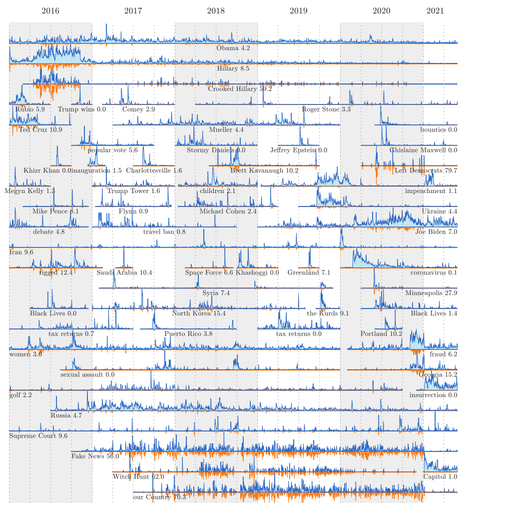
Emotional turbulence:



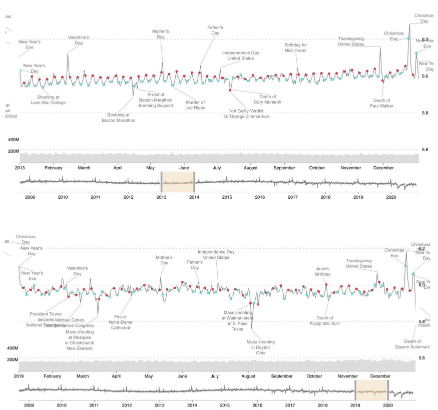
http://hedonometer.org/



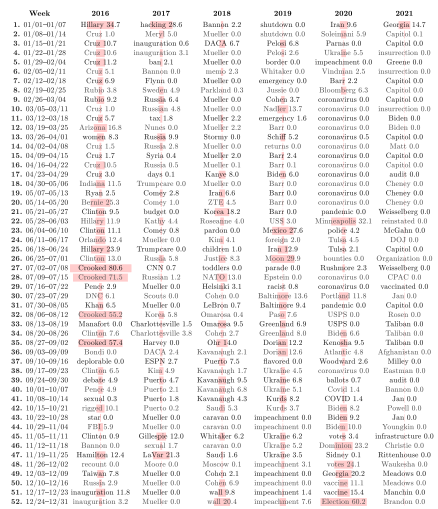
http://compstorylab.org/trumpstoryturbulence/

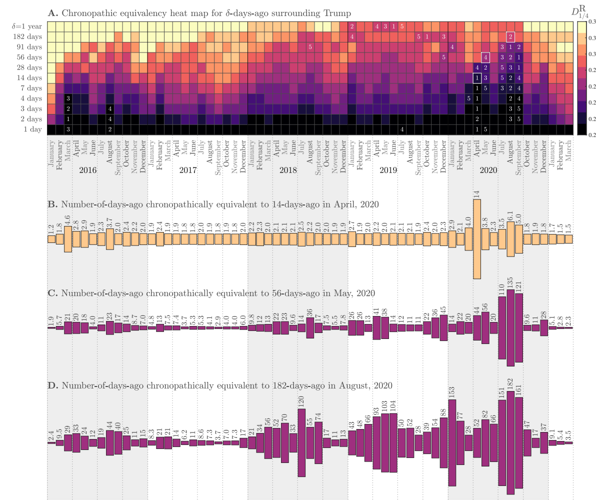


Emotional turbulence:

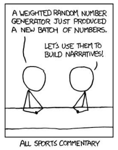


http://hedonometer.org/

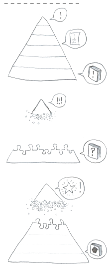




Understanding the Sociotechnocene—Stories:



xkcd.com/904/



- Toward a Science of Stories.
- Claim: Homo narrativus—we run on stories.
- "What's the John Dory?"
- "They've lost the plot/thread"
- Narrative hierarchies and scalability of stories.
- Research: Real-time and offline extraction of metaphors, frames, plots, narratives, conspiracy theories, and stories from large-scale text.
- Research: The taxonomy of human stories.
- To be built: Storyscopes—improvable, online, interactive instruments.

ding!

On Instagram at pratchett_the_cat

References I

[1] P. E. Auerswald.
The Code Economy: A Forty-Thousand Year History.
Oxford University Press, 2017.

[2] B. Boyd.
On the Origin of Stories: Evolution, Cognition, and Fiction.
Belknap Press, 2010.

[3] 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

References II

[4] E. Cheng.
How to bake pi: An edible exploration of the mathematics of mathematics.
Basic Books, 2015.

[5] B. Christian and T. Griffiths.
Algorithms to Live By.
Macmillan, 2016.

[6] 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

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

References III

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

[9] 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 IV

[10] P. S. Dodds, J. R. Minot, M. V. Arnold, T. Alshaabi, J. L. Adams, D. R. Dewhurst, A. J. Reagan, and C. M. Danforth.
Fame and Ultrafame: Measuring and comparing daily levels of 'being talked about' for United States' presidents, their rivals, God, countries, and K-pop, 2019.
Available online at https://arxiv.org/abs/1910.00149. pdf

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

References V

Available online at https://arxiv.org/abs/2008.13078. pdf

[12] P. S. Dodds, J. R. Minot, M. V. Arnold, T. Alshaabi, J. L. Adams, A. J. Reagan, and C. M. Danforth.
Computational timeline reconstruction of the stories surrounding Trump: Story turbulence, narrative control, and collective chronopathy, 2020.
https://arxiv.org/abs/2008.07301. pdf

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

References VI

[14] M. Erwig.
Once Upon an Algorithm.
MIT Press, 2017.

[15] J. Gottschall.
The Storytelling Animal: How Stories Make Us Human.
Mariner Books, 2013.

[16] 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

[17] E. Katz and P. F. Lazarsfeld.
Personal Influence.
The Free Press, New York, 1955.

References VII

[18] T. Kuran.
Now out of never: The element of surprise in the east european revolution of 1989.
World Politics, 44:7–48, 1991. pdf

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

[20] 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 VIII

[21] 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

[22] J. R. Minot, M. V. Arnold, T. Alshaabi, C. M. Danforth, and P. S. Dodds.
Ratioing the President: An exploration of public engagement with Obama and Trump on Twitter, 2020.
Available online at https://arxiv.org/abs/2006.03526. pdf

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

References IX

[24] 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.

[25] 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

[26] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery.
Numerical Recipes in C.
Cambridge University Press, second edition, 1992.

References X

[27] M. Puchner.
The Written World: How Literature Shaped Civilization.
Random, 2017.

[28] M. J. Salganik, P. S. Dodds, and D. J. Watts.
An experimental study of inequality and unpredictability in an artificial cultural market.
Science, 311:854–856, 2006. pdf

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

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

References XI

[31] 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.

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

[33] D. J. Watts and P. S. Dodds.
Influentials, networks, and public opinion formation.
Journal of Consumer Research, 34:441–458, 2007. pdf

References XII

[34] 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