Visualization-Guided Text Analysis in Python

Jason Kessler

Presentation for Indie Thinkers

June 22, 2021

Overview

- If you'd like to follow along, please install Scattertext and related packages
  - https://github.com/jasonkessler/IndiethinkersTutorial

- Visualize variation in language
  - Between groups (often by using proxies for membership)
    - Left vs. right
    - Recipients of the Pfizer or Moderna vaccines
      - By gender
  - Over time
- We will first discuss some work in visualizing categorized text
- Then discuss how to use Scattertext to analyze social media chatter around Vaccines
- Discuss work in visualizing language change over time
- Demo a new feature of Scattertext which allows you to see how topics evolve over time

OKCupid: an online dating site

Which words and phrases statistically distinguish ethnic groups and genders?

hobos

almond butter

Bikram yoga

100 Years of Solitude

Christian Rudder. “The Real ‘Stuff White People Like’” 2010. https://www.gwern.net/docs/psychology/okcupid/therealstuffwhitepeoplelike.html
Rudder was OKCupid’s founding data scientist. This blog post has since been removed from okcupid.com.

OKCupid: Words and phrases that distinguish white men.

In general, I won't comment too much on these lists, because the whole point of this piece is to let the groups speak for themselves, but I have to say that the mind of the white man is the world's greatest sausagefest. Unless you're counting **Queens of the Stone Age**, there is not even one vaguely feminine thing on his list, and as far as broad categories go we have: sweaty guitar rock, bro-on-bro comedies, things with engines, and dystopias.

Christian Rudder. “The Real ‘Stuff White People Like’” 2010. https://www.gwern.net/docs/psychology/okcupid/therealstuffwhitepeoplelike.html
Rudder was OKCupid’s founding data scientist. This blog post has since been removed from okcupid.com.

OKCupid: Words and phrases that distinguish Latin men.

Explanation

Music and dancing—**merengue**, **bachata**, **reggaeton**, **salsa**—are obviously very important to Latinos of both genders. The men have two other fascinating things going on: an interest in telling you about their sense of humor (**i'm a funny guy**, **very funny**, **outgoing and funny**, etc.) and an interest in industrial strength ass-kicking (**mma**, **ufc**, **boxing**, **marines**, etc.) Basically, if a Latin dude tells you a joke, you should laugh.

Christian Rudder. “The Real ‘Stuff White People Like’” 2010. https://www.gwern.net/docs/psychology/okcupid/therealstuffwhitepeoplelike.html
Rudder was OKCupid’s founding data scientist. This blog post has since been removed from okcupid.com.

Word Use and Gender in Facebook Statuses

The good:

- Word clouds force you to hunt for the most impactful terms
- You end up examining the long tail in the process
- Compactly represent a lot of phrases and topics

Lyle Ungar
2013 AAAI
Tutorial

Word Use and Gender in Facebook Statuses

a a a
correlation strength

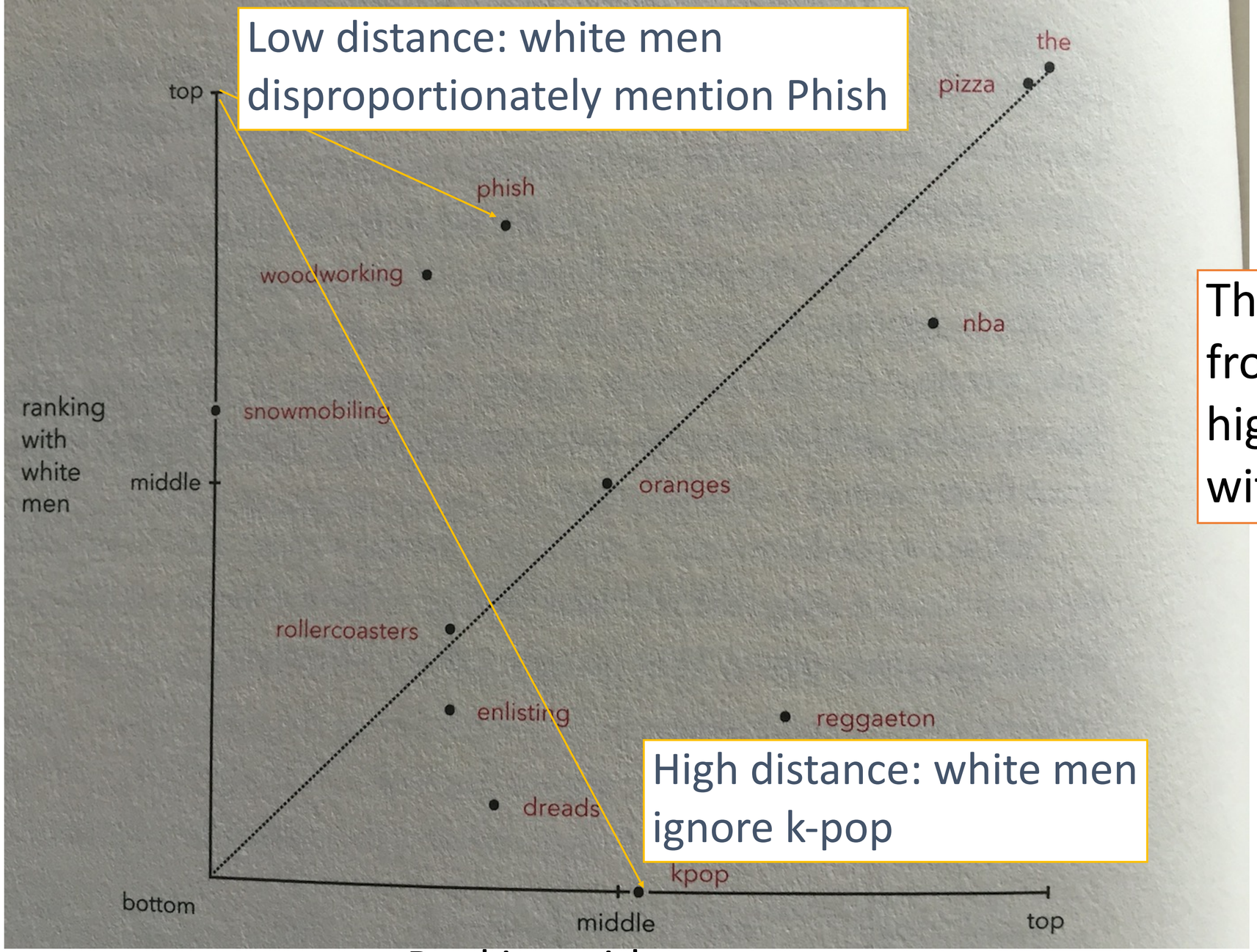
relative frequency

b b b
prevalence in topic

The bad:

- "Mullets of the Internet" --Jeffrey Zeldman, **2005**
- Longer phrases are are more prominent.
- Ranking is unclear
- Does size indicate higher frequency?

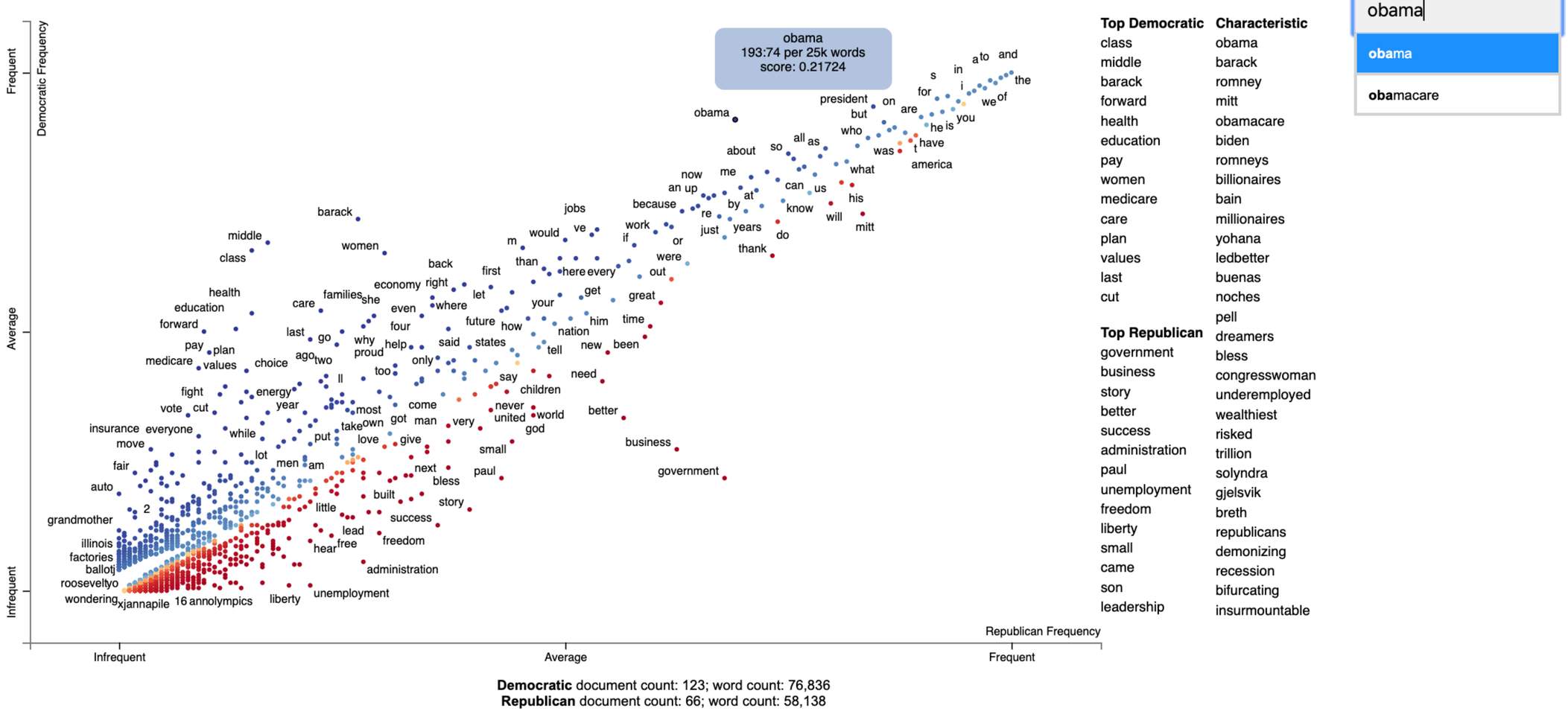
Lyle Ungar
2013 AAAI
Tutorial



The smaller the distance from the top left, the higher the association with white men.

Source: Christian Rudder. Dataclysm. 2014.

Source: Christian Rudder. Dataclysm. 2014.



Top Democratic	Characteristic
class	obama
middle	barack
barack	romney
forward	mitt
health	obamacare
education	biden
pay	romneys
women	billionaires
medicare	bain
care	millionaires
plan	yohana
values	ledbetter
last	buenas
cut	noches
	pell
Top Republican	dreamers
government	bless
business	congresswoman
story	underemployed
better	wealthiest
success	risked
administration	trillion
paul	solyndra
unemployment	gjelsvik
freedom	breth
liberty	republicans
small	demonizing
came	recession
son	bifurcating
leadership	insurmountable

obama

obama

obamacare

Democratic document count: 123; word count: 76,836
Republican document count: 66; word count: 58,138

Term: obama

Democratic frequency:
193 per 25,000 terms
919 per 1,000 docs
Some of the 535 mentions:

RICHARD DURBIN

It was a cold, cold January afternoon when Barack **Obama** lifted his hand from Abraham Lincoln's Bible and looked out on an America facing an economic collapse.

Well, President **Obama** and millions of American families think it's a great idea for America.

And with President **Obama** and Vice President Joe Biden in the White House, we will.

Republican frequency:
74 per 25,000 terms
606 per 1,000 docs
Some of the 167 mentions:

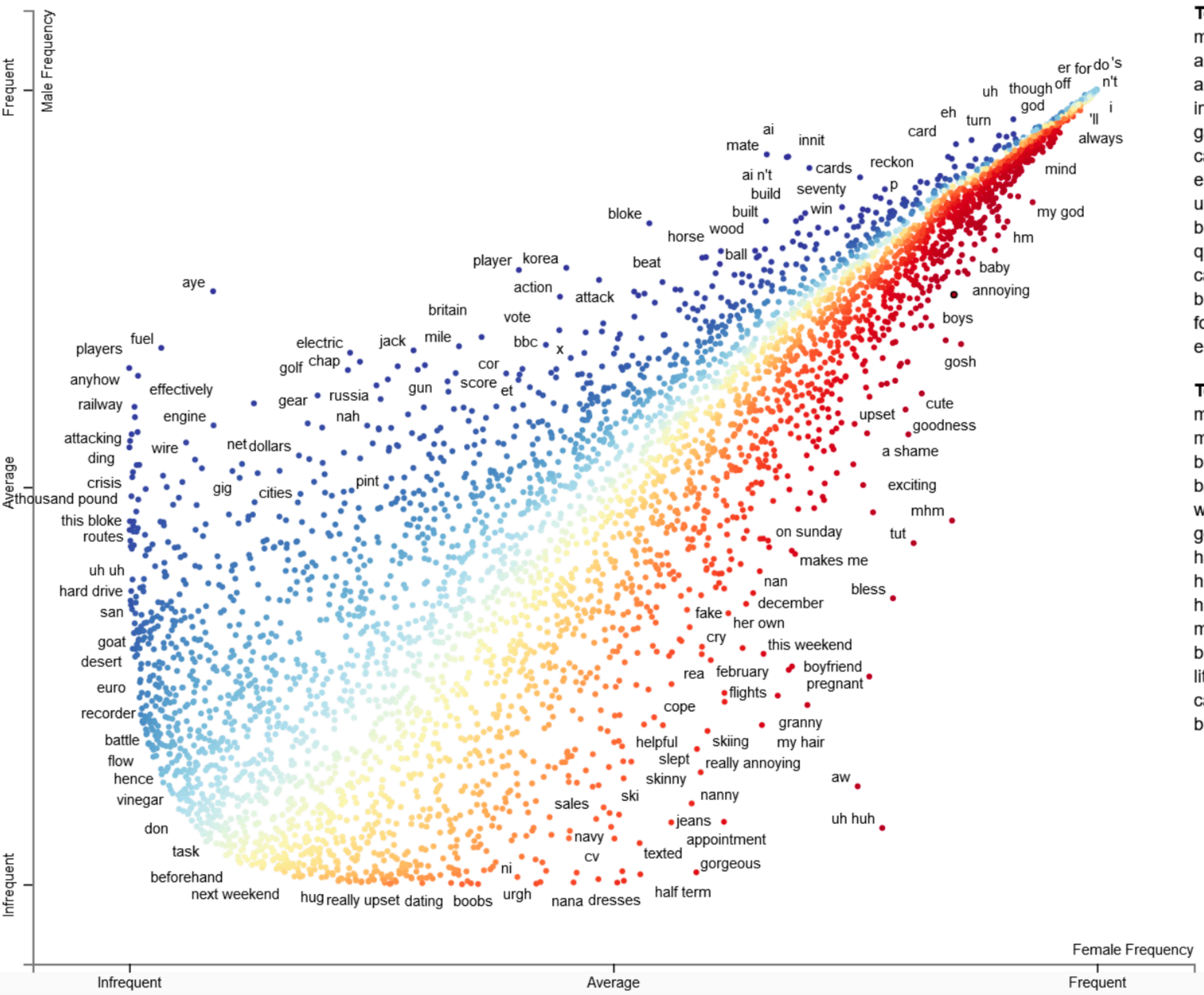
MITT ROMNEY

I wish President **Obama** had succeeded because I want America to succeed.

If you felt that excitement when you voted for Barack **Obama,** shouldn't you feel that way now that he's President Obama?

Some of the companies we helped start are names you — you know and you've heard from tonight: an office company called Staples, where I'm pleased to see the **Obama** campaign's been shopping — — today Steel

Jason S. Kessler. Scattertext: a Browser-Based Tool for Visualizing how Corpora Differ. ACL System Demonstrations. Vancouver, BC. 2017.



Top Male

mate
ai n't
ai
innit
game
cards
eh
uh
bloke
quid
card
built
football
em

Top Female

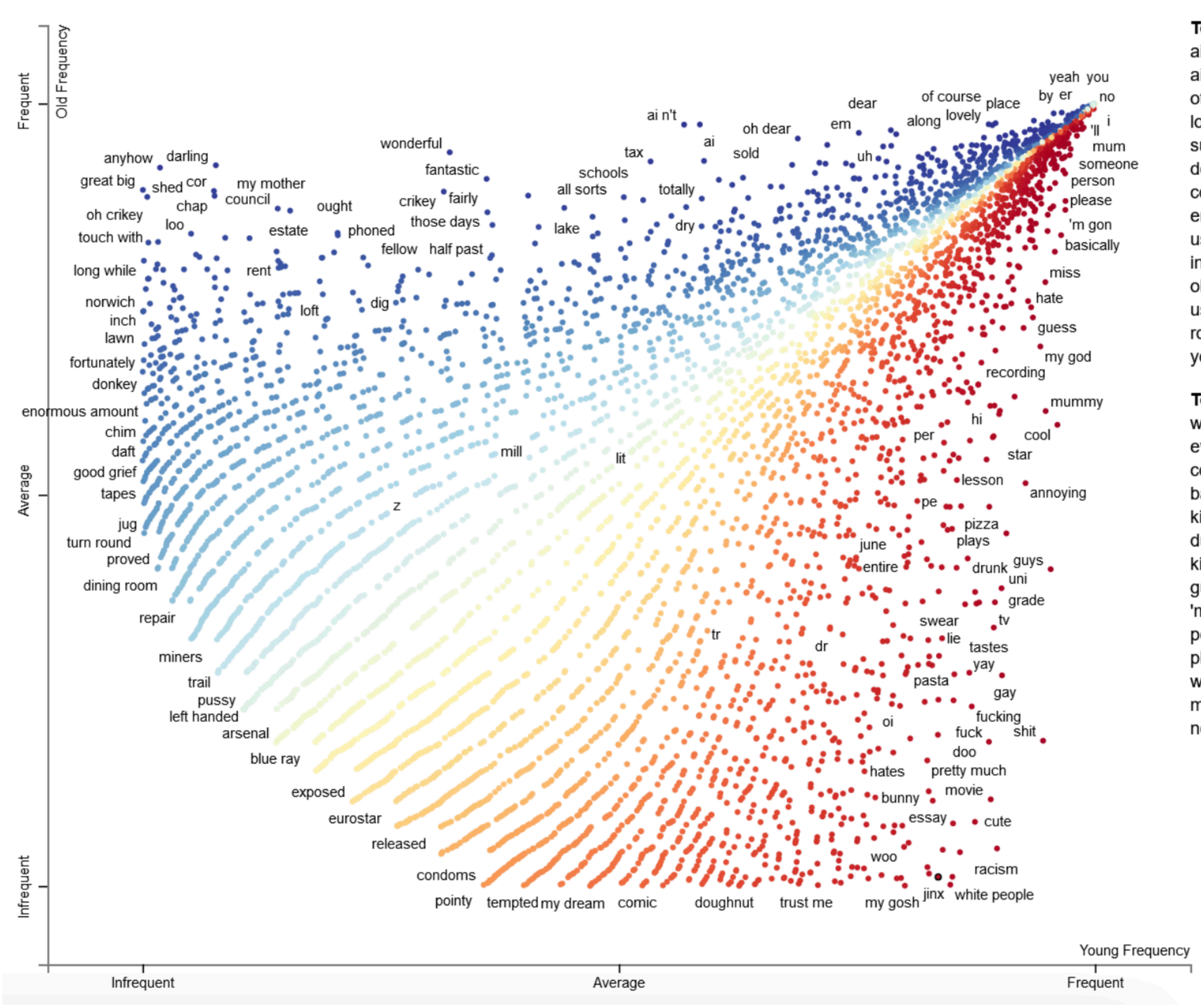
my god
mhm
blah
baby
weekend
gosh
hair
huh
hm
mm hm
birthday
literally
cake
blah blah

Scattertext in the wild:

- Identifying differences between gender and age in SpokenBNC2014
  - Differences in "minimal particles" are particularly interesting
    - M: Eh, um, em,
    - F: mhm, hum, hm, mm hm

Plotting roughly on frequency rank; ties are broken alphabetically

Andreas Liesenfeld, Gabor Parti, Yu-Yin Hsu, Chu-Ren Huang. Predicting gender and age categories in English conversations using lexical, non-lexical, and turn-taking features. Pacling 2020.



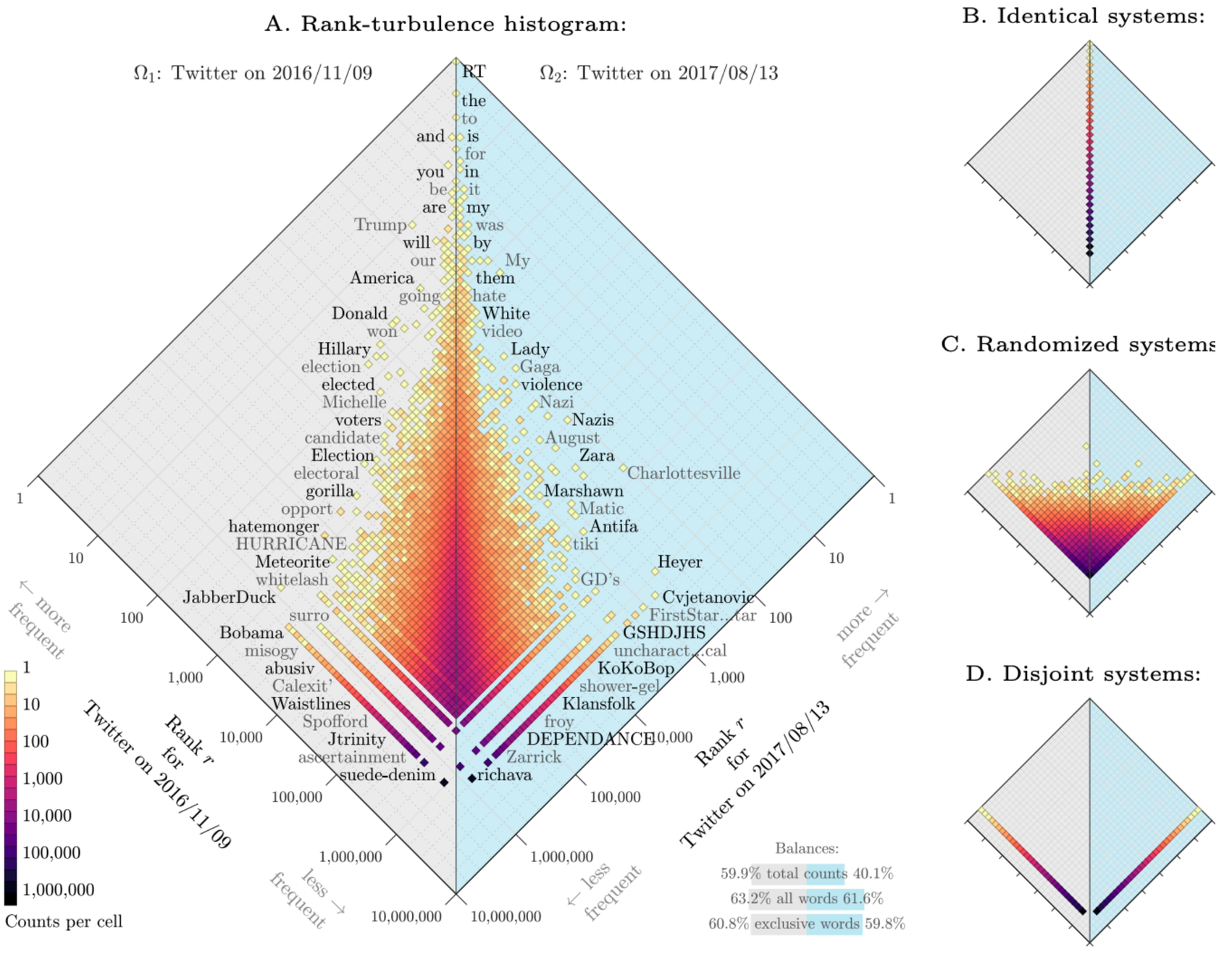
Top Old
ai
ai n't
of course
lovely
suppose
dear
course
em
used to
innit
oh dear
used
road
years

Top Young
wanna
everyone
cool
basically
kind of
dunno
kind
guys
'm gon
person
play
wait
maybe
need

Scattertext in the wild: -Old (>=70) vs. young (<=19)

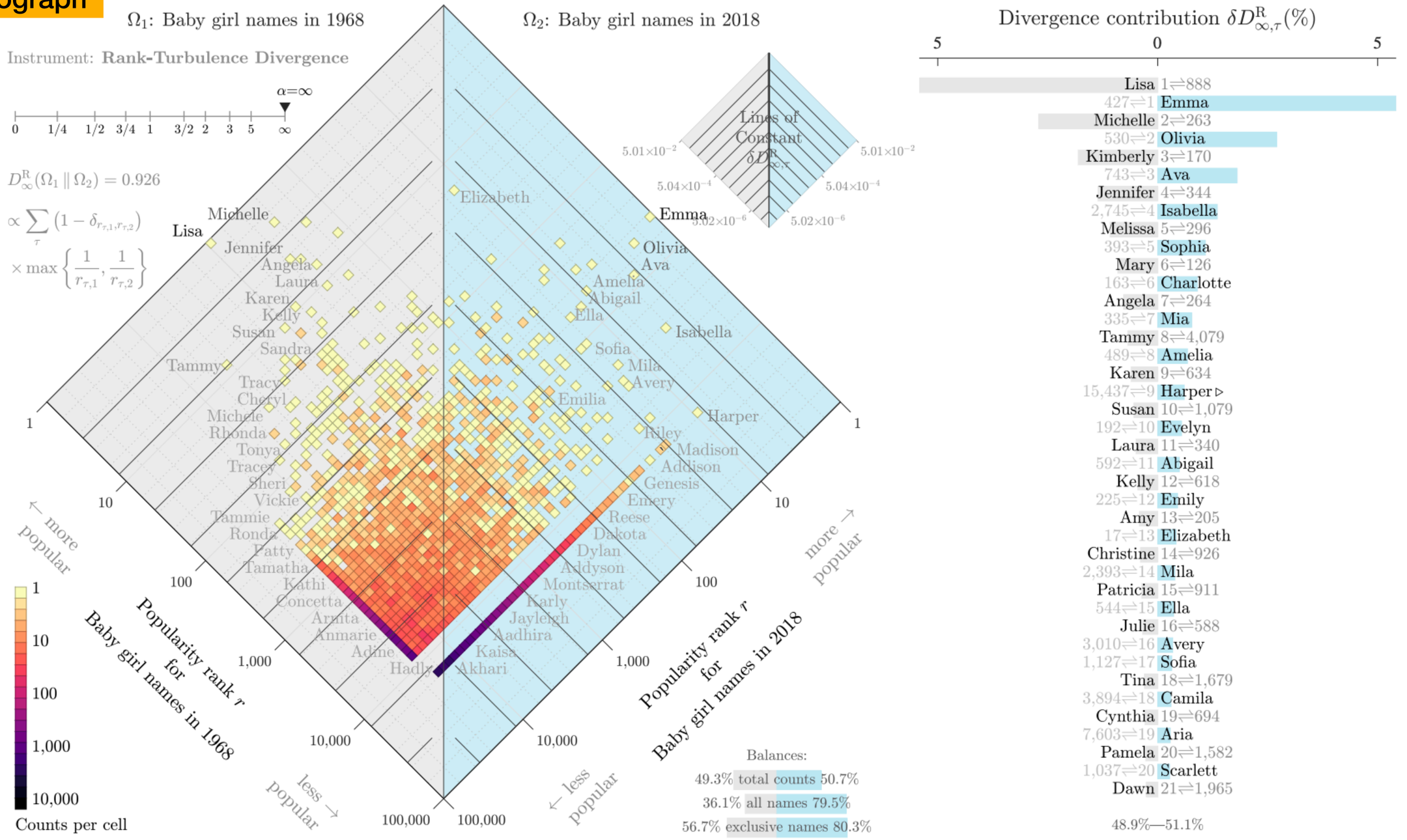
Author's note (Oct 2020): statement on the use of social categories in this study This work involves the labelling of participants for social categories related to gender and age. We caution against the use of this heuristic due to the risk of promoting a biased view on the topics. We would like to encourage those interested in the computational modelling of social categories to join the discussion on these concerns and consider participating in efforts to build more inclusive resources for the study of the topics.

Andreas Liesenfeld, Gabor Parti, Yu-Yin Hsu, Chu-Ren Huang. Predicting gender and age categories in English conversations using lexical, non-lexical, and turn-taking features. Pacling 2020.



Dodds, P. et al. “Allotaxonometry and rank-turbulence divergence: A universal instrument for comparing complex systems.” arXiv: Physics and Society (2020): n. pag.

“Allotaxonograph”



Dodds, P. et al. “Allotaxonometry and rank-turbulence divergence: A universal instrument for comparing complex systems.” arXiv: Physics and Society (2020)

Using Scattertext Part 1

- What can we learn from how people tweet about Moderna and Pfizer?
- Jupyter notebook demo
  - https://nbviewer.jupyter.org/github/JasonKessler/IndiethinkersTutorial/blob/main/Moderna%20vs.%20Pfizer%20Analysis.ipynb
- This includes enough code for you to easily replicate the study
  - And explore your own corpoa

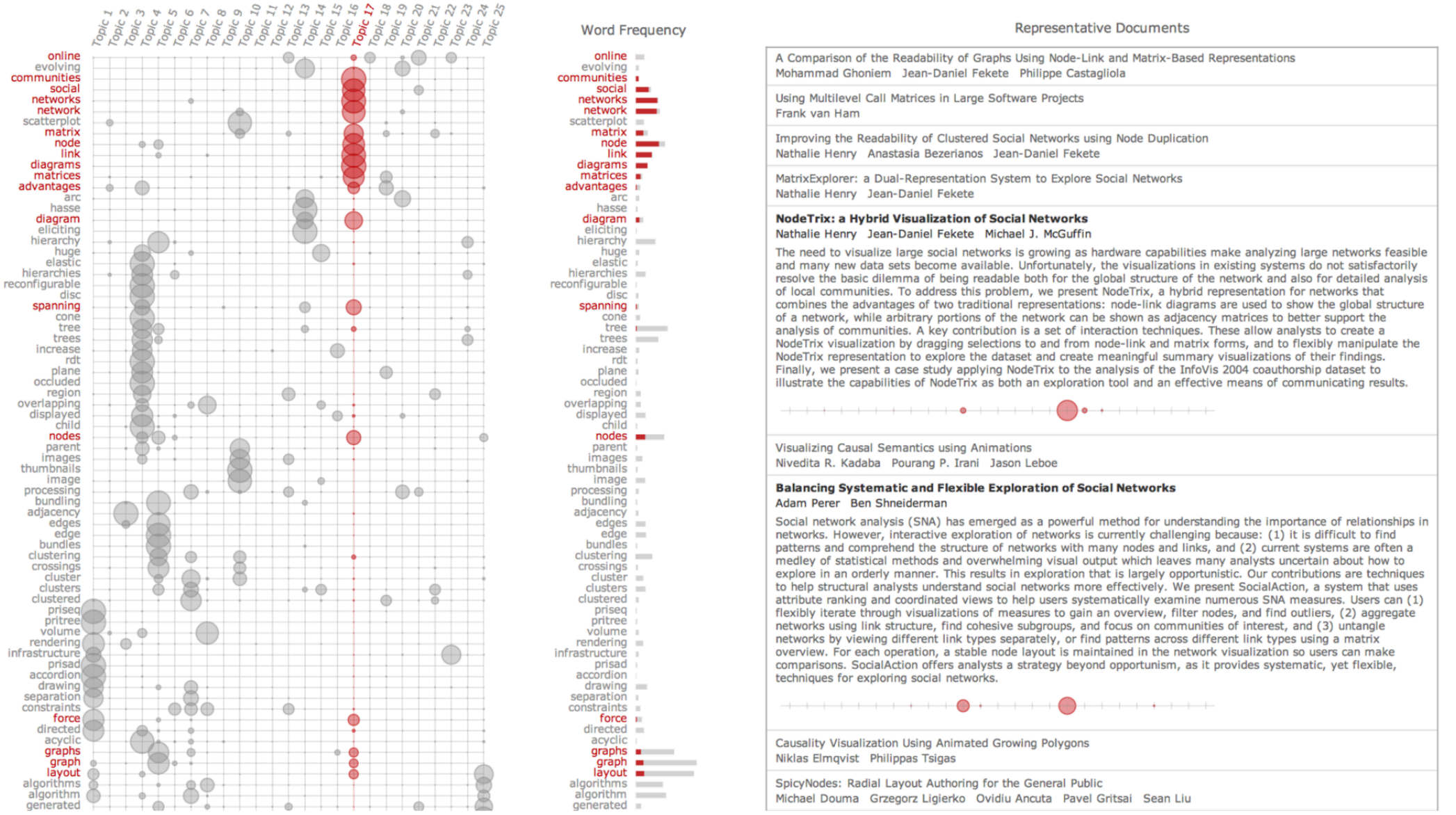


Figure 3: The Termite system. When a topic is selected in the term-topic matrix (left), the systems visualizes the word frequency distribution relative to the full corpus (middle) and shows the most representative documents (right).

Jason Chuang, Christopher D. Manning, Jeffrey Heer. Termite: Visualization Techniques for Assessing Textual Topic Models. Advanced Visual Interfaces, 2012

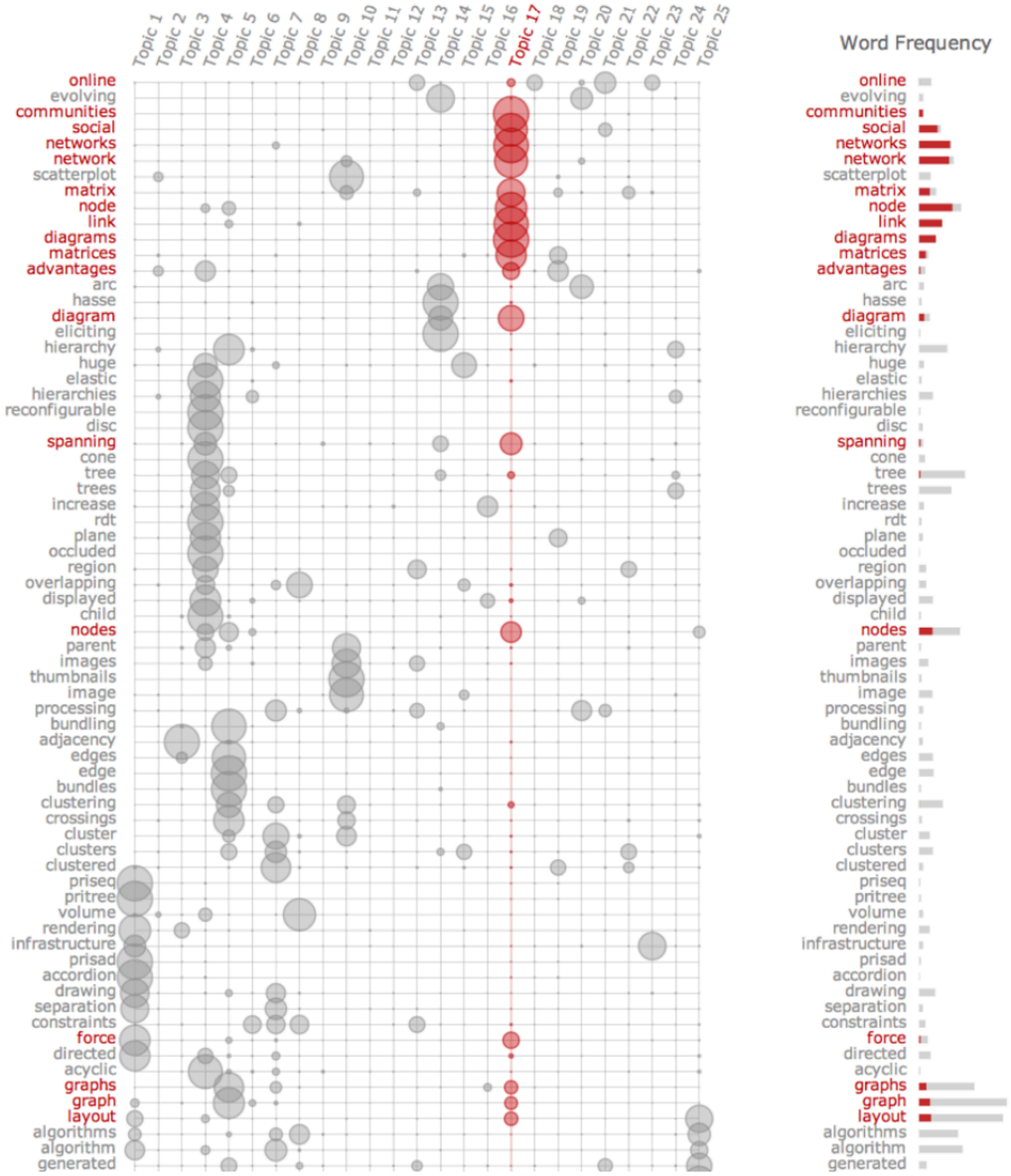


Figure 3: The Termite system. When a topic is selected ir word frequency distribution relative to the full corpus (midd

Words are included based on “saliency”

- 10-250 terms shown; monitor dependent
- Distinctiveness
  - Sum KL-divergence(P(topic|word)||P(topic))

$$distinctiveness(w) = \sum_T P(T|w) \log \frac{P(T|w)}{P(T)}$$

$$saliency(w) = P(w) \times distinctiveness(w)$$

KLD is used for circle sizes

Ordered using “seriation”

- If word B closely follows word A in the list, “A B” is likely to be a phrase
- Neighboring words tend to contain phrases in order
- Online, evolving, communities, social, networks, network
- G^2 for collocation; Bond Energy Algorithm for ordering

Jason Chuang, Christopher D. Manning, Jeffrey Heer. Termite: Visualization Techniques for Assessing Textual Topic Models. Advanced Visual Interfaces, 2012

C. Collins, F. B. Viégas, and M. Wattenberg, "Parallel Tag Clouds to Explore and Analyze Facted Text Corpora," *VAST* 2009.

Visual Encoding

- One column per *Circuit*
- Top *N* words only (scrollable)

- Text Size → significance score
- Order → alphabetic

Significance score:

Based on G^2:
- Similar to a chi-square contingency test

Size indicates significance score, top K most significant terms are included

Order is alphabetical, which keeps the same tags at similar horizontal levels
- Position changes could falsely imply difference in significance

Difference in position is more noticeable than difference in size

Can’t directly indicate change in freq.

C. Collins, F. B. Viégas, and M. Wattenberg, “Parallel Tag Clouds to Explore and Analyze Facted Text Corpora,” *VAST* 2009.

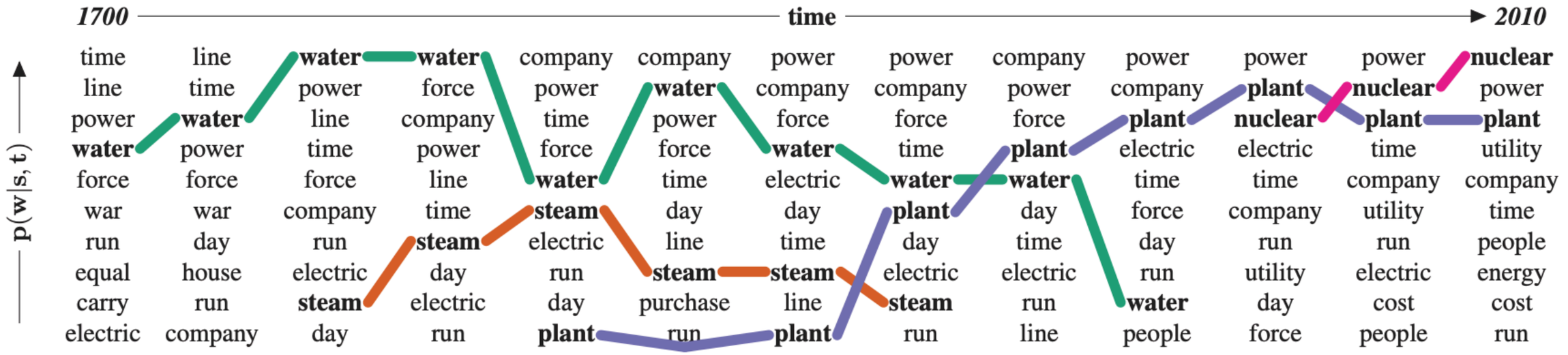


Figure 5: Sense-internal temporal dynamics for the *energy* sense of the word `power` (violet/6 in Figure 4). Columns show the ten most highly associated words for each time interval for the period between 1700 and 2010 (ordered by decreasing probability). We highlight how four terms characteristic of the sense develop over time (see {water, steam, plant, nuclear} in the figure).

Lea Frermann and Mirella Lapata. A Bayesian Model of Diachronic Meaning Change. TACL 2016.

Figure 1: Two-dimensional visualization of semantic change in English using SGNS vectors.[2] **a**, The word *gay* shifted from meaning "cheerful" or "frolicsome" to referring to homosexuality. **b**, In the early 20th century *broadcast* referred to "casting out seeds"; with the rise of television and radio its meaning shifted to "transmitting signals". **c**, *Awful* underwent a process of pejoration, as it shifted from meaning "full of awe" to meaning "terrible or appalling" (Simpson et al., 1989).

William L. Hamilton, Jure Leskovec, Dan Jurafsky. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. ACL 2016.

Using Scattertext Part 2

- What can we learn about the evolution of the terms “woke” and “BIPOC” on Twitter?
  - Early demo of Scattertext’s newest feature: diachronic tables
  - Inspired by Parallel Tag Clouds
    - Columns are sorted by l1-regression coefficients, sizes are proportional to frequency
    - Also includes residual dispersion scatterplot. Novel measure of dispersion.
- Jupyter notebook demo
  - https://nbviewer.jupyter.org/github/JasonKessler/IndiethinkersTutorial/blob/main/Evolutions%20of%20Woke%20and%20BIPOC.ipynb

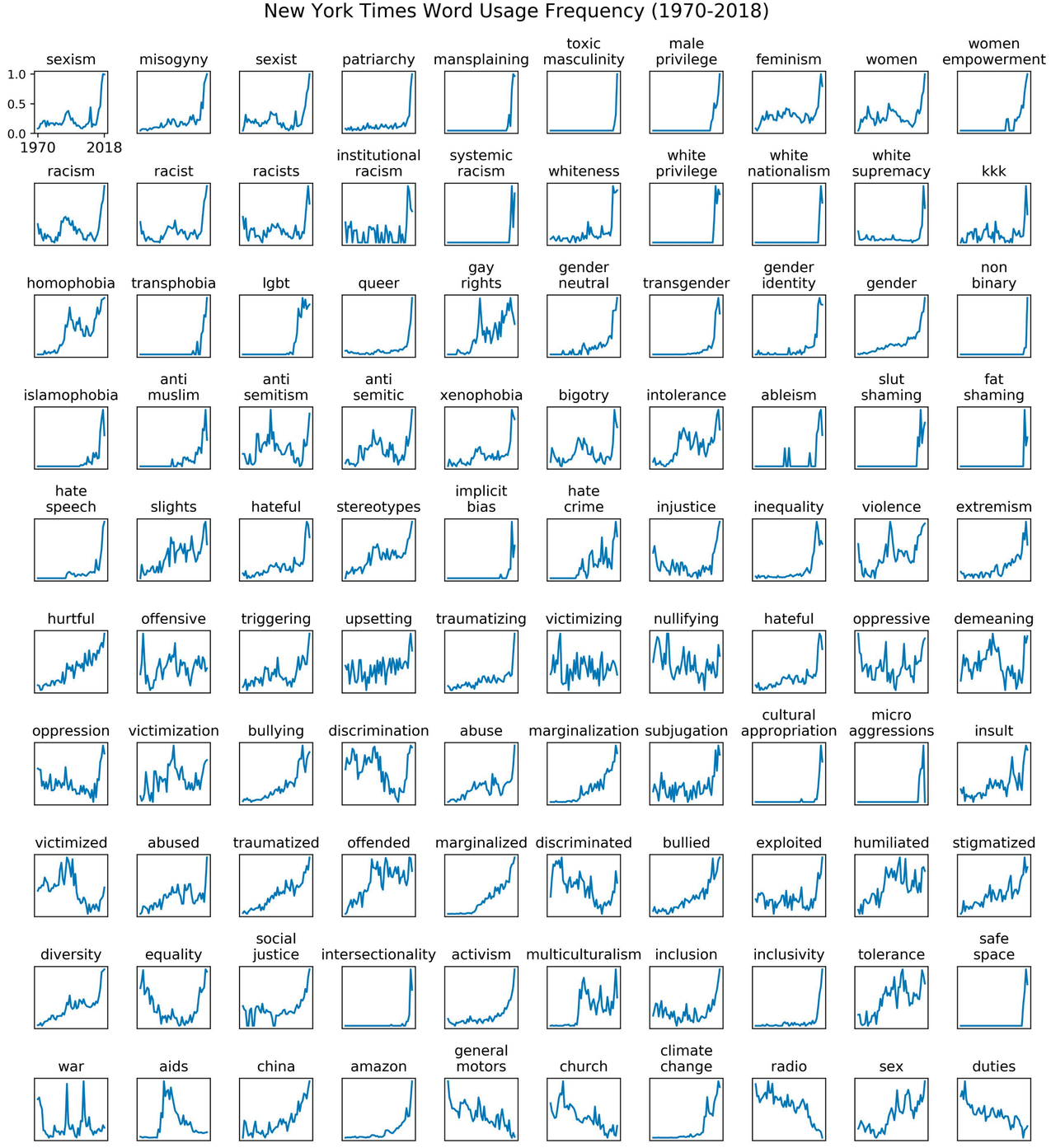
SparkClouds

Users studies show this outperforms Parallel Tag Clouds

Terms not used in epoch being examined are blurred. Size is relative to frequency at epoch.

Issues: number of terms is limited; most frequent terms aren’t always the most informative

Bongshin Lee Amy Karlson Nathalie Henry Riche. SparkClouds: Visualizing Trends in Tag Clouds. In IEEE TVCG (Proceedings of Infovis 2010) | November 2010



https://twitter.com/DavidRozado/status/1134041329292460032