Measuring Cultural Affinity and Influence from Aggregated Diffusion of Language Change

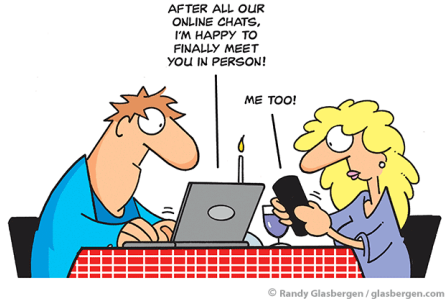
Jacob Eisenstein
@jacobeisenstein

Georgia Institute of Technology

February 14, 2015

What would it take for you to use a word that you had never used before?

The rise of digital writing



- Informal communication is increasingly conducted in digital, written form.
- This shift is driving fundamental changes in the nature of written language.

The consequences.

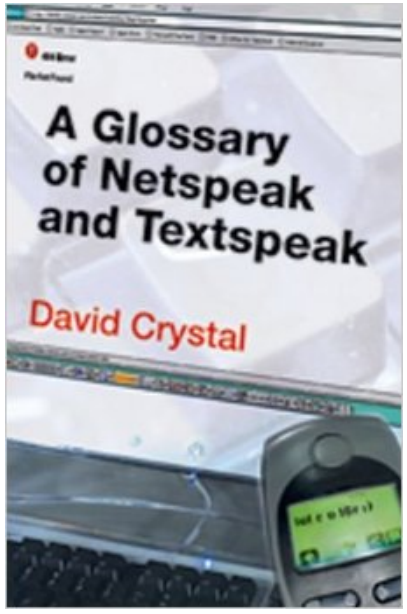
...dats why pluto is pluto it can neva b a star

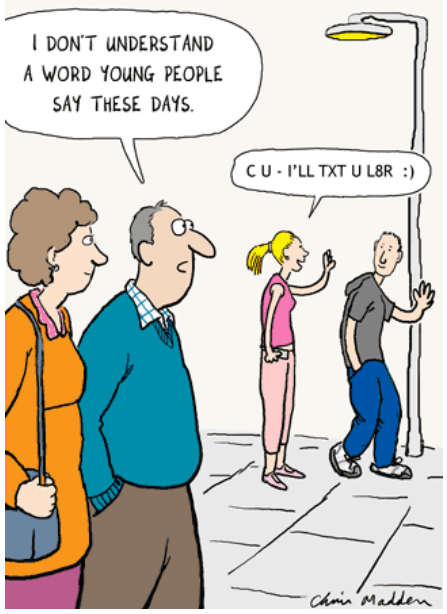
Work on farm Fri. Burning piles of brush WindyFire got out of control. Thank God for good naber He help get undr control PantsBurnLegWound.

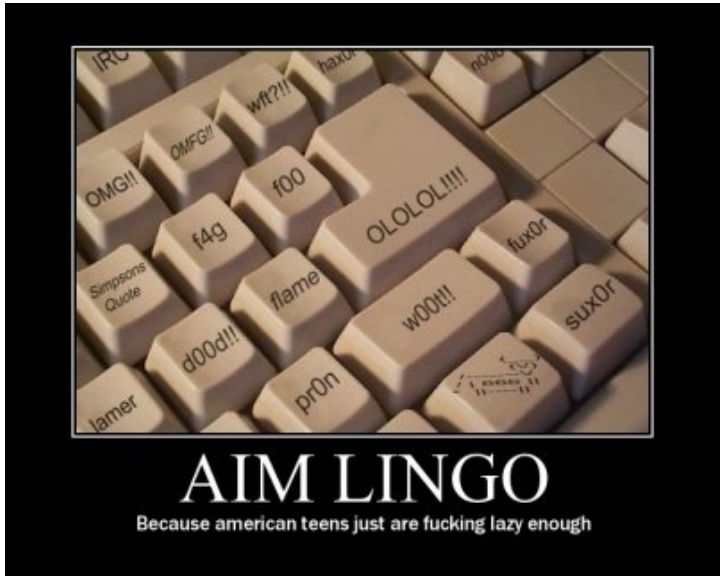
Boom! Ya ur website suxx bro

michelle obama great. job. and. whit all my. respect she. look. great. congrats. to. her.

A New Netspeak Variety?



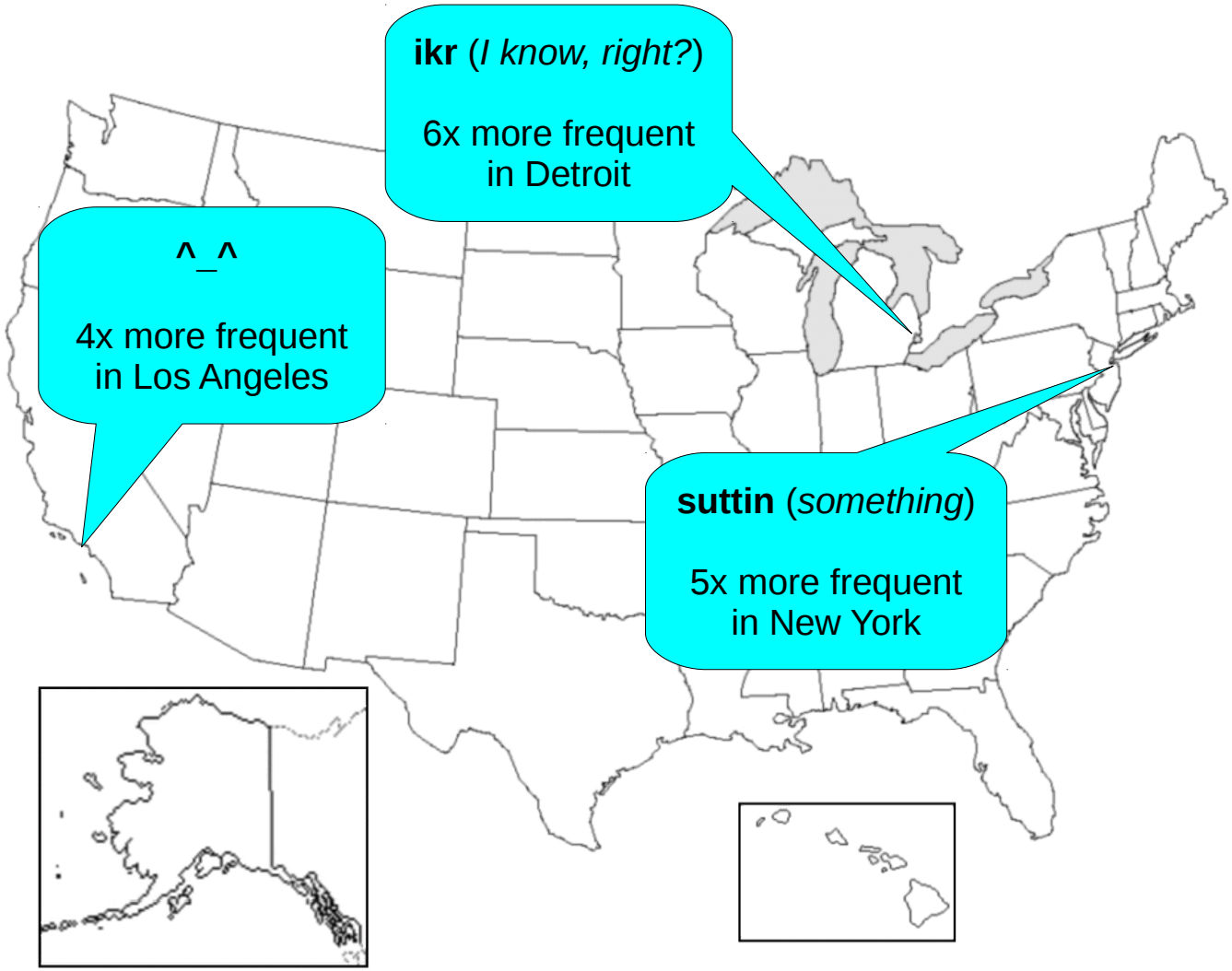




- Some have heralded the birth of a new "netspeak" variety.
- But netspeak is remarkably diverse, reflecting geography, demographics, politics, and more.

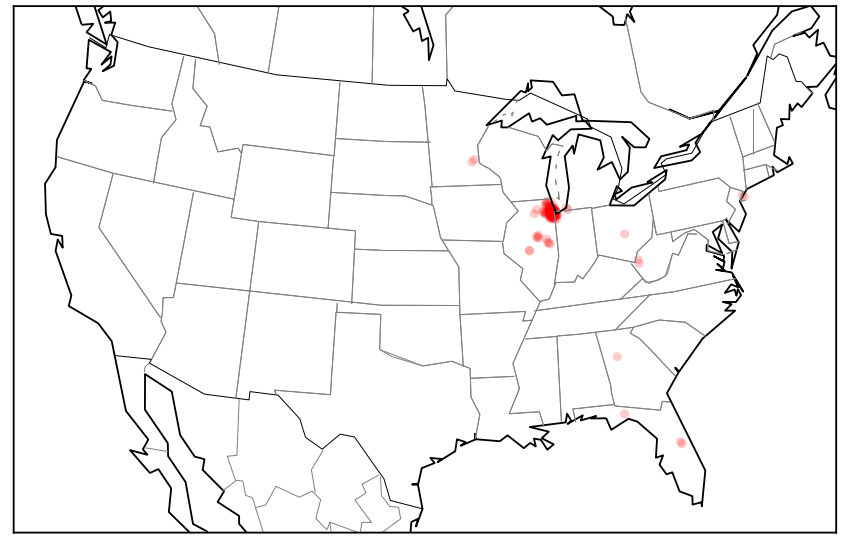
Not netspeak, but netspeak**S**

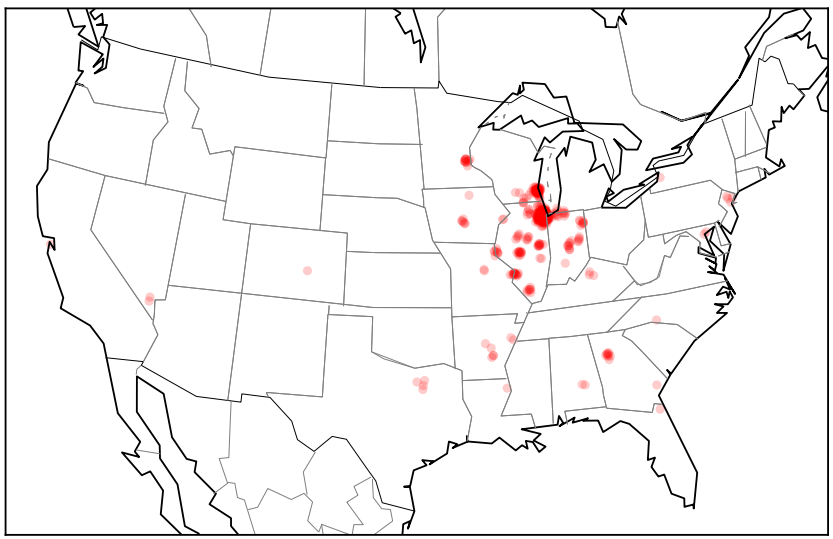
On Twitter:

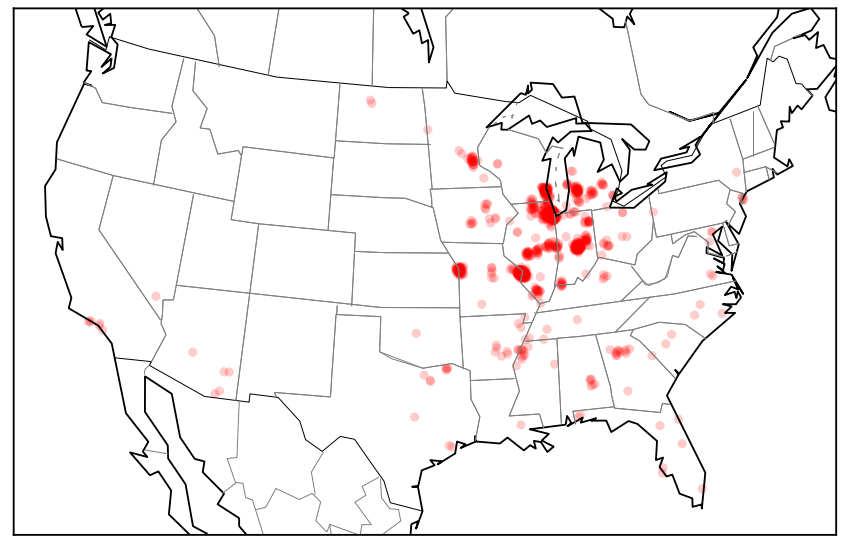


Change from 2010-2012: lbvs

tell ur momma 2 buy me a car lbvs

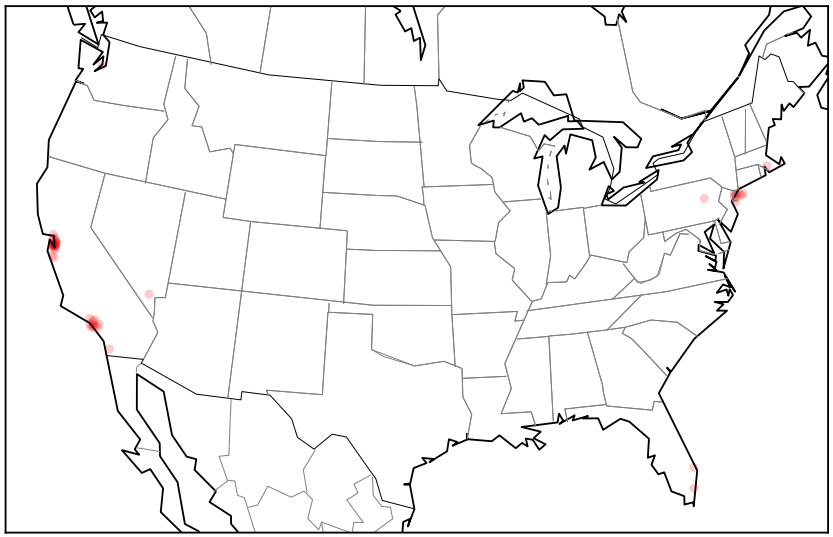


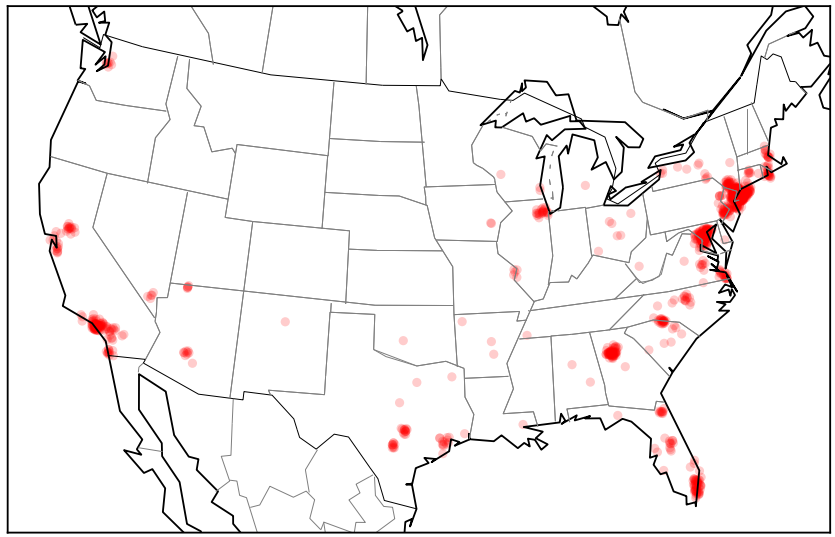


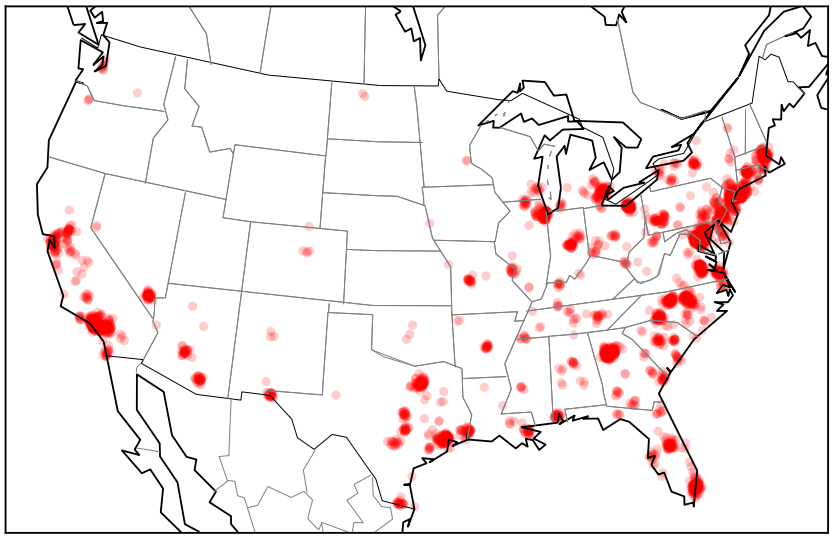


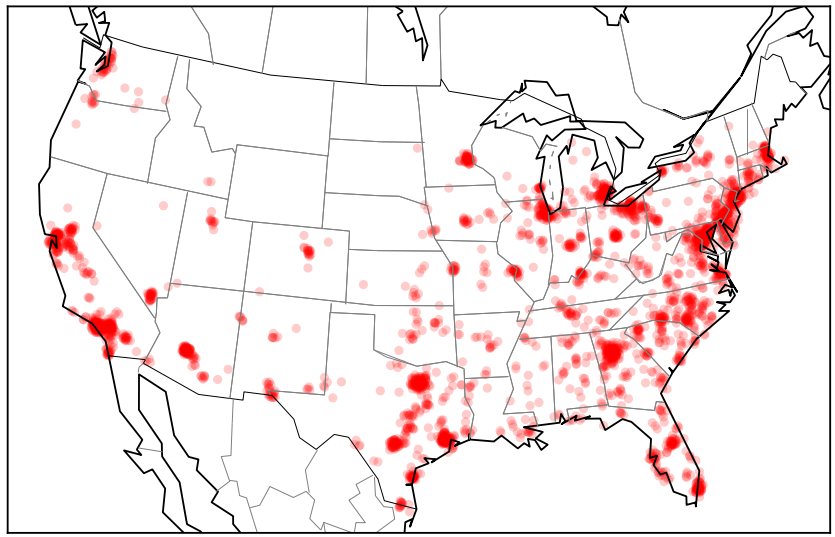
Change from 2009-2012: -__-

flight delayed -__- just what i need



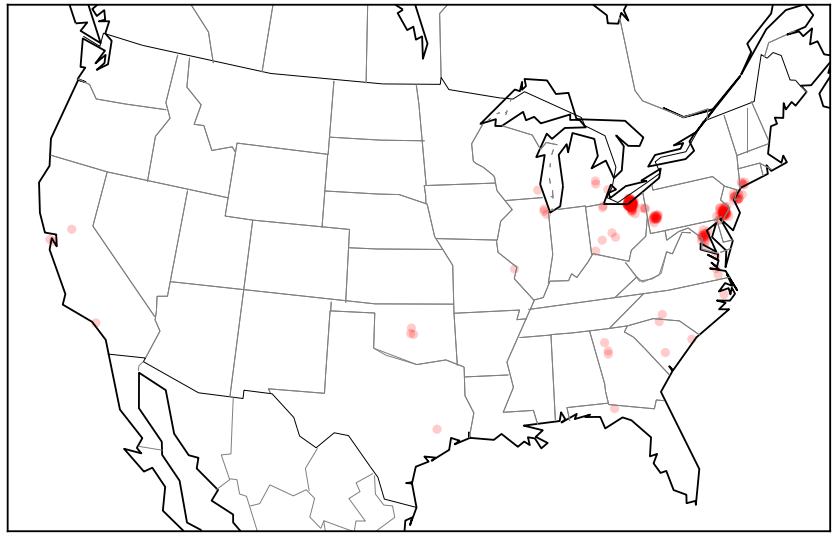


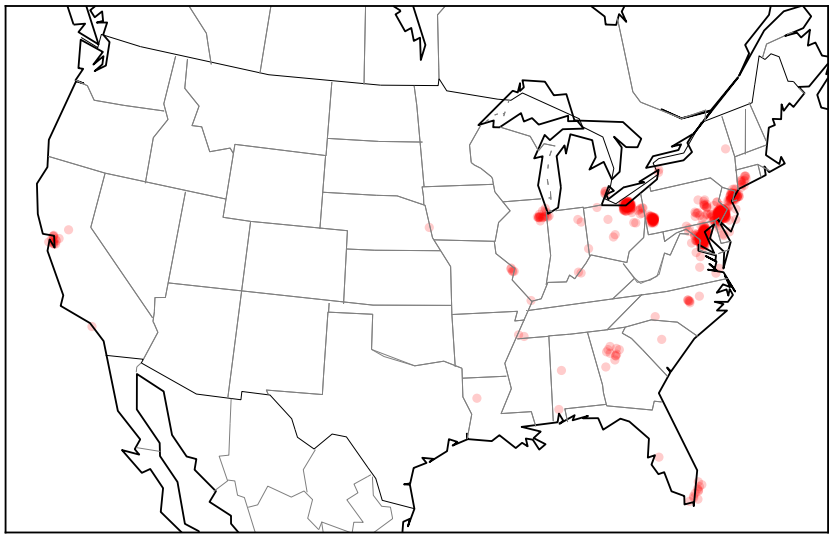


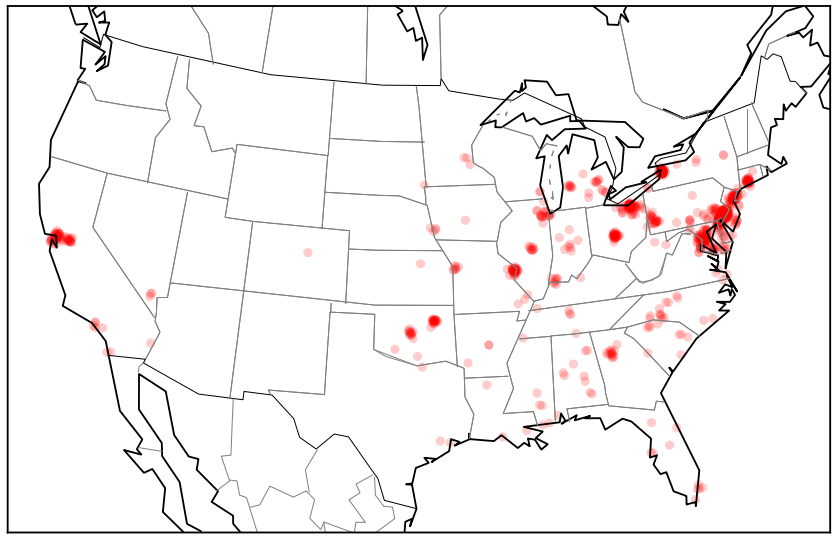


Change from 2009-2012: ctfu

@name lmao! haahhaa ctfu!



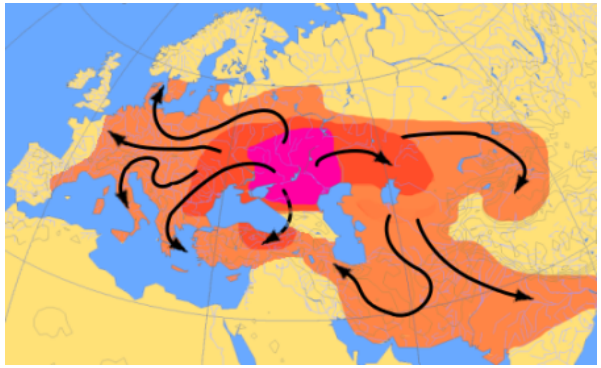




Models of language change

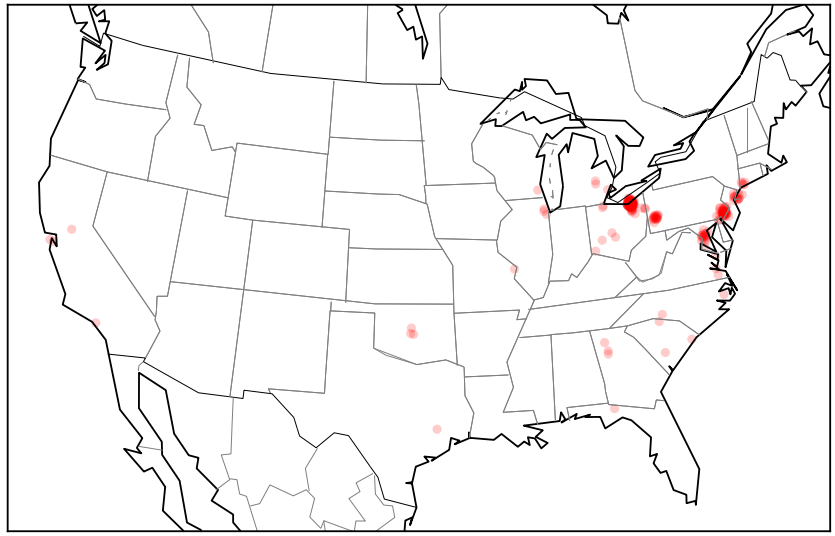
Wave model change over space

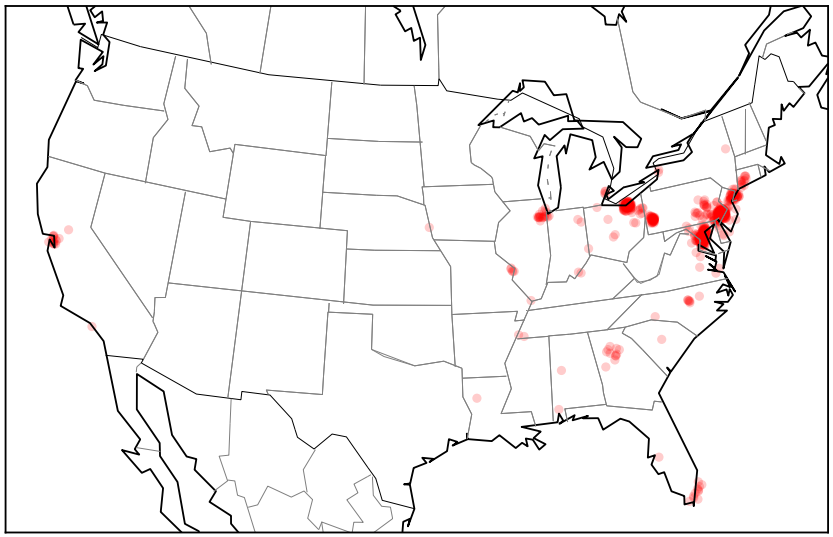
Cascade model big cities first

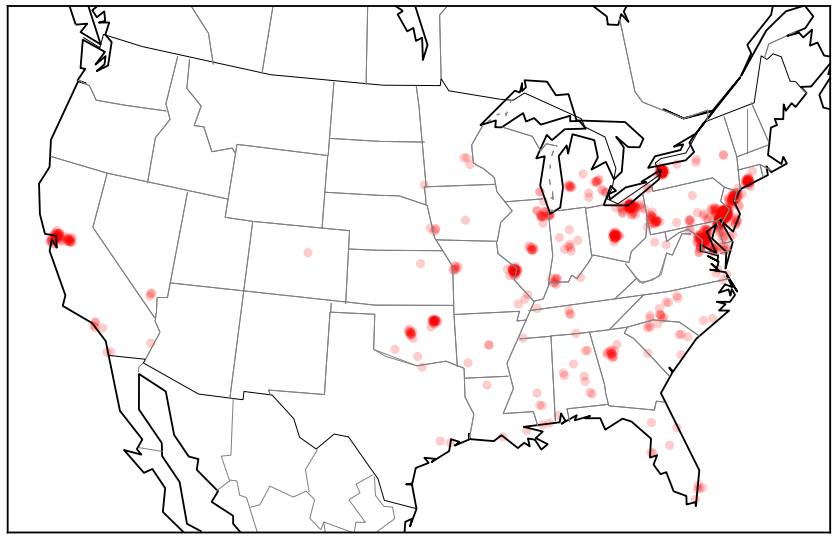


Change from 2009-2012: ctfu

@name lmao! haahhaa ctfu!



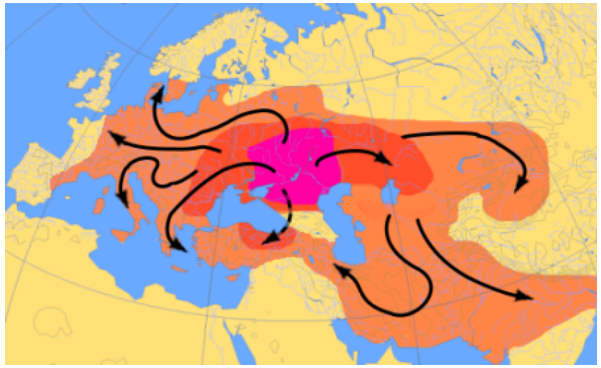




Models of language change

Wave model change over space

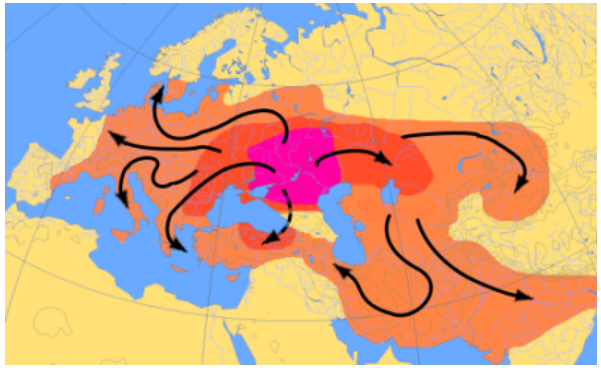
Cascade model big cities first



Models of language change

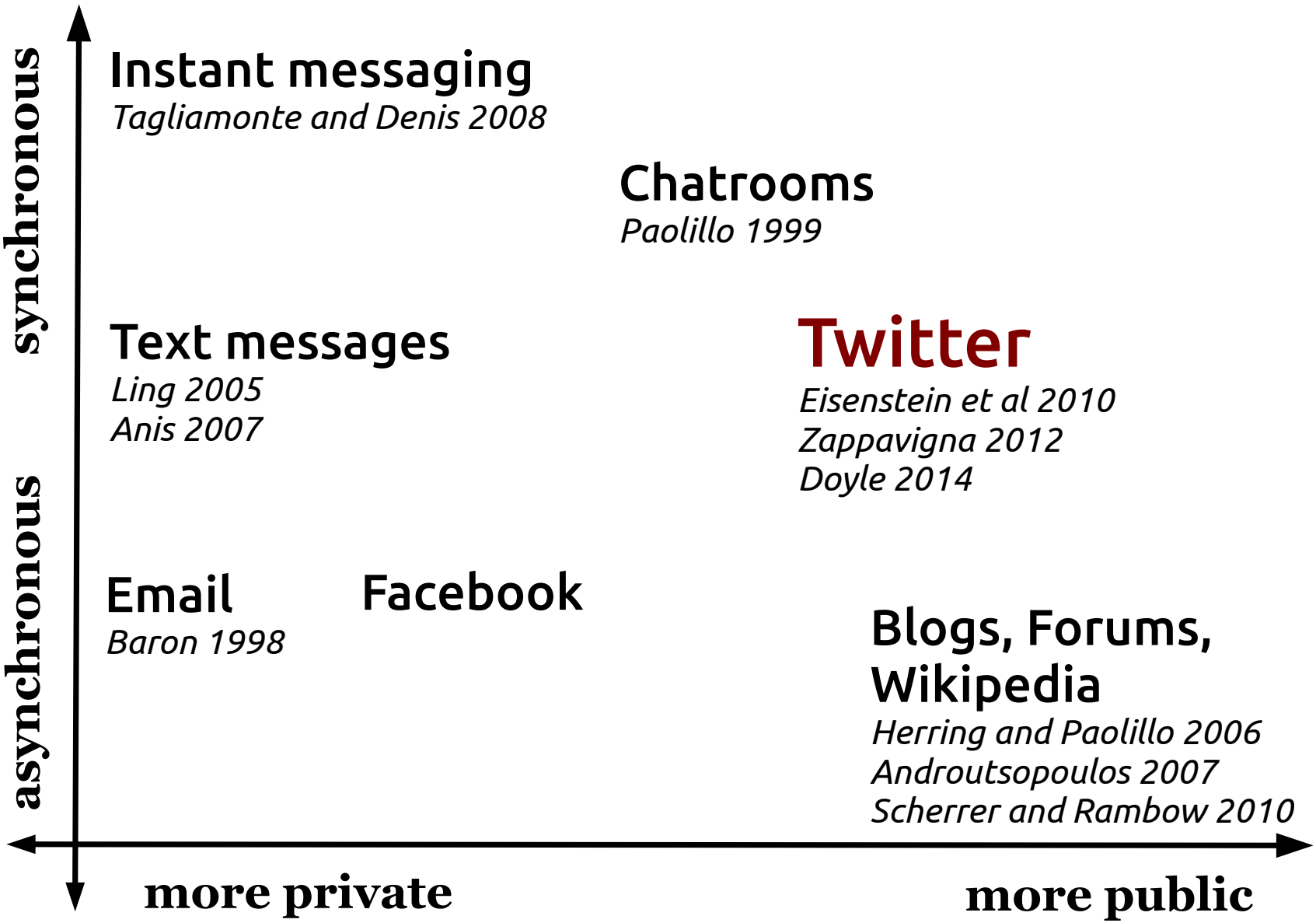
Wave model change over space

Cascade model big cities first



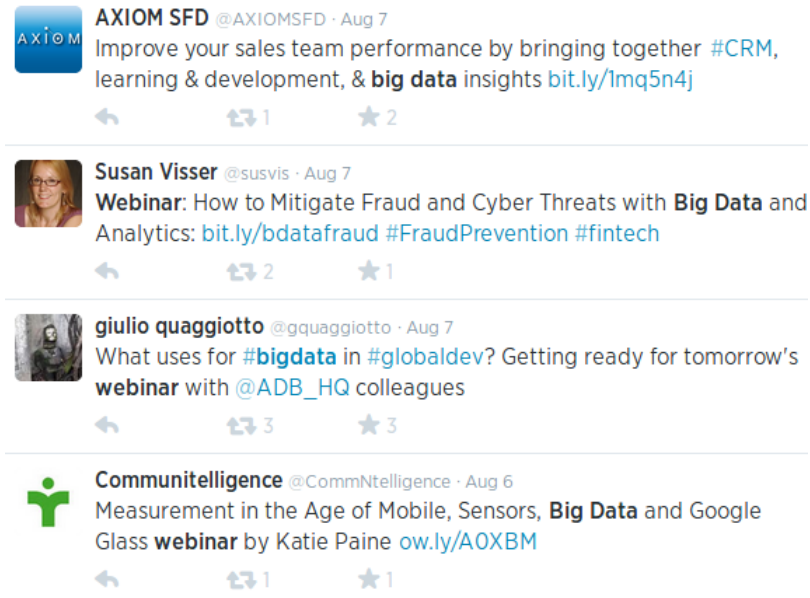
- Language change is modulated by sociocultural factors, like **power**, **affinity**, and **influence**.
- With large-scale records of digital writing, it is now possible to **quantify** the interplay of geography, demographics, and culture.

A landscape of digital communication



Twitter

- 140-character messages
- Each user has a custom **timeline** of people they've chosen to **follow**.
- Most data is publicly accessible, along with social network and geographical metadata.



Who are these people?

	2013	2014
All internet users	*18%*	*23%**
Men	17	24*
Women	18	21
White, Non-Hispanic	16	21 *
Black, Non-Hispanic	29	27
Hispanic	16	25
18-29	31	37
30-49	19	25
50-64	9	12
65+	5	10*
High school grad or less	17	16
Some college	18	24
College+ (n=685)	18	30*
Less than $30,000/yr	17	20
$30,000-$49,999	18	21
$50,000-$74,999	15	27*
$75,000+	19	27*
Urban	18	25*
Suburban	19	23
Rural	11	17

(Pew Research Center)

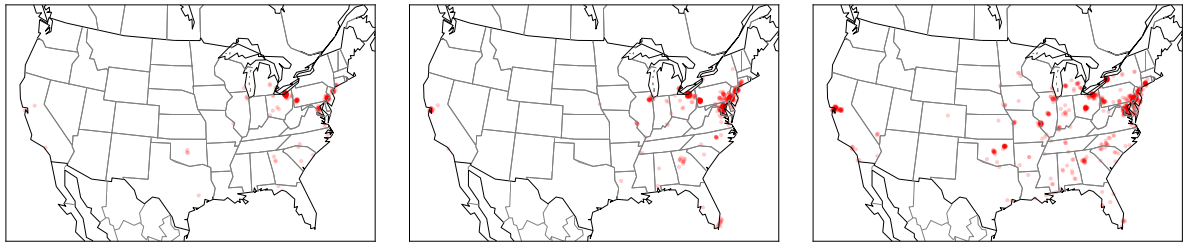
- % of online adults who use Twitter; per-message statistics will differ.
- Representativeness concerns are real, but there are potential solutions.
- Social media has important representativeness advantages too.

Dataset

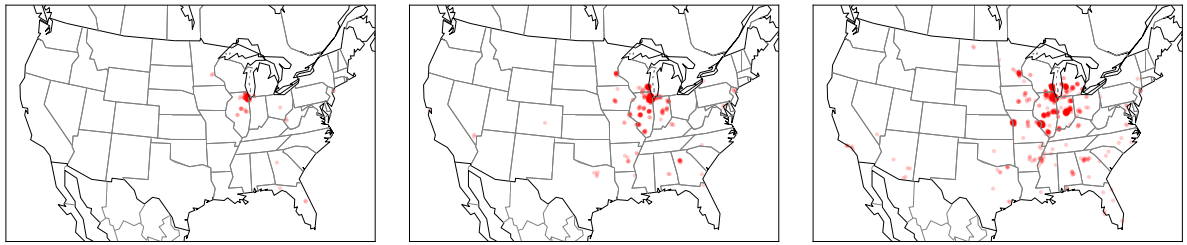
- 114 million messages, all geolocated in the United States using cellphone GPS.
- August 2009 to September 2012, aggregated by week
- 2.77 million user accounts, aggregated into the 200 largest metropolitan areas in the USA.
- Filters
  - Messages: no retweets, URLs
  - Users: no celebrities
  - Words: no hashtags, usernames

An aggregate model of lexical diffusion

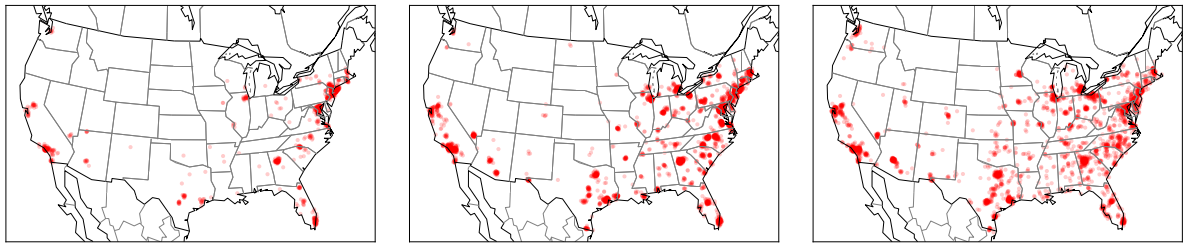
ctfu



lbvs

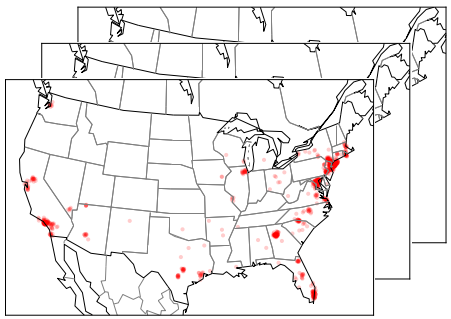


-__-

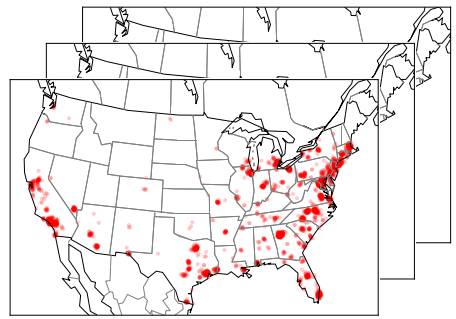


- Thousands of words have changing frequencies.
- Each spatiotemporal trajectory is idiosyncratic.
- What's the aggregate picture?

Language change as an autoregressive process



$\eta_2 \sim N(A\eta_1, \Sigma)$



$\eta_3 \sim N(A\eta_2, \Sigma)$

$$
\begin{aligned}
c_{\text{ctfu},1} &\sim \text{Binomial}(f(\eta_{\text{ctfu},1}), N_1) \\
c_{\text{hella},1} &\sim \text{Binomial}(f(\eta_{\text{hella},1}), N_1) \\
&\dots
\end{aligned}
$$

$$
\begin{aligned}
c_{\text{ctfu},2} &\sim \text{Binomial}(f(\eta_{\text{ctfu},2}), N_2) \\
c_{\text{hella},2} &\sim \text{Binomial}(f(\eta_{\text{hella},2}), N_2) \\
&\dots
\end{aligned}
$$

Estimating parameters of this autoregressive process reveals geographic pathways of diffusion across thousands of words.

Aggregating across words

- Let a represent region-to-region linguistic "influence"

  $a_{i,j}$ is the influence between region i and region j.

- Let c represent the word counts per region

  $c_{w,r,t}$ is the count of word w in region r at time t.

- We want the **maximum likelihood estimate**,

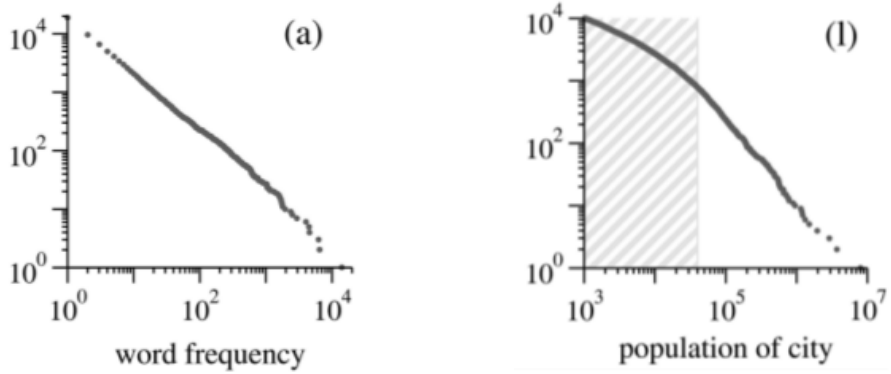
$$\hat{a} = \arg\max_{a} P(c; a)$$

But first we have to define this probability.

Why raw word counts won't work

We observe counts $c_{w,r,t}$ for word w in region r at time t. How does $c_{w,r,t}$ influence $c_{w,r',t+1}$?

- Both word counts and city sizes follow power law distributions, with lots of zero counts.



- Exogenous events such as pop culture and weather introduce global temporal effects.
- Twitter's sampling rate is inconsistent, both spatially and temporally.

Latent activation model

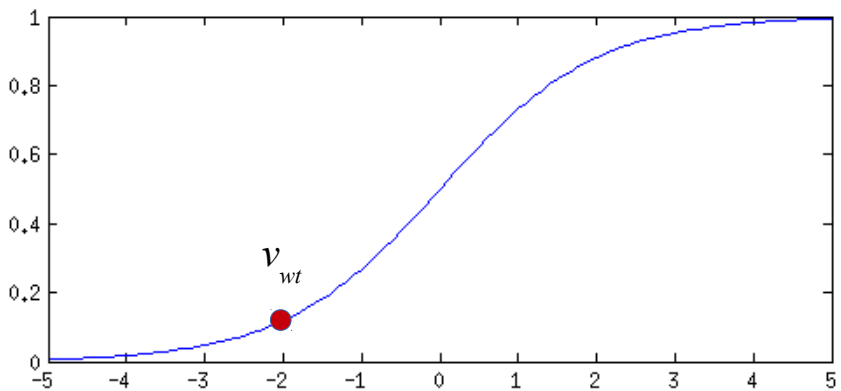
$$c_{w,r,t} \sim \text{Binomial}(\beta_{w,r,t}, s_{r,t})$$

Latent activation model

$$c_{w,r,t} \sim \text{Binomial}(\beta_{w,r,t}, s_{r,t})$$
$$\beta_{w,r,t} = \text{Logistic}(\nu_{w,t} + \mu_{r,t} + \eta_{w,r,t})$$

Latent activation model

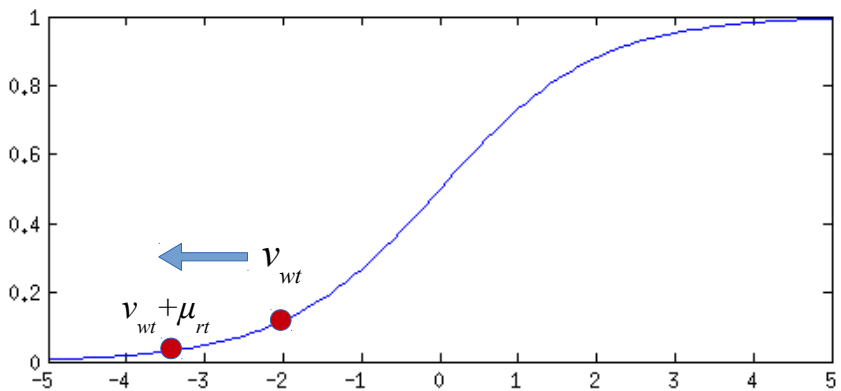
$$c_{w,r,t} \sim \text{Binomial}(\beta_{w,r,t}, s_{r,t})$$
$$\beta_{w,r,t} = \text{Logistic}(\nu_{w,t} + \mu_{r,t} + \eta_{w,r,t})$$



- Base word log-probability

Latent activation model

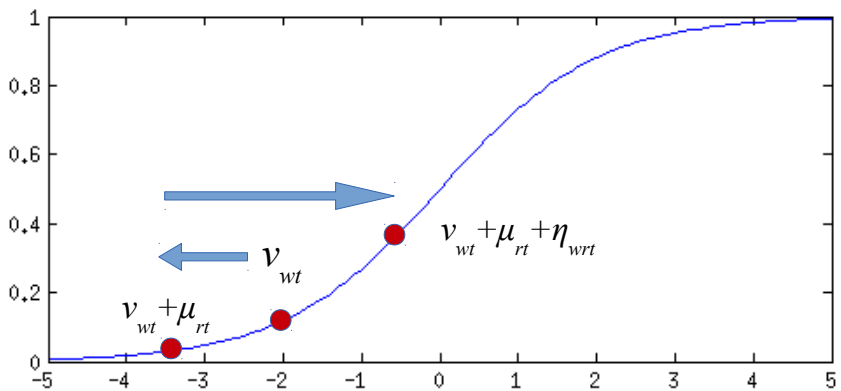
$$c_{w,r,t} \sim \text{Binomial}(\beta_{w,r,t}, s_{r,t})$$
$$\beta_{w,r,t} = \text{Logistic}(\nu_{w,t} + \mu_{r,t} + \eta_{w,r,t})$$



- Base word log-probability
- City-specific "verbosity"

Latent activation model

$$c_{w,r,t} \sim \text{Binomial}(\beta_{w,r,t}, s_{r,t})$$
$$\beta_{w,r,t} = \text{Logistic}(\nu_{w,t} + \mu_{r,t} + \eta_{w,r,t})$$



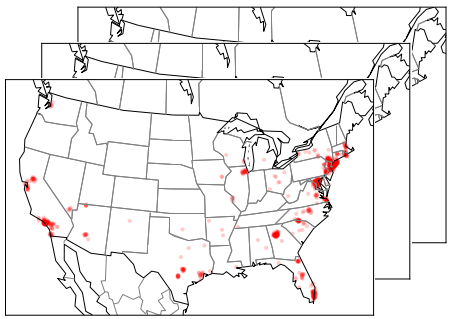
- Base word log-probability
- City-specific "verbosity"
- **Spatio-temporal activation**

Dynamics model

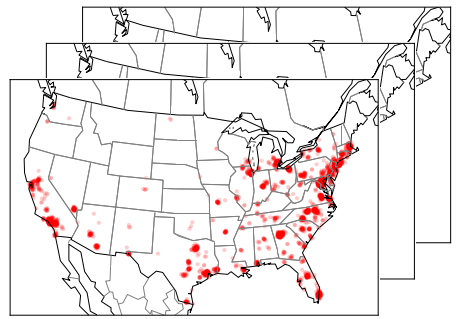
$$
\begin{aligned}
c_{w,r,t} &\sim \text{Binomial}(\beta_{w,r,t}, s_{r,t}) \\
\beta_{w,r,t} &= \text{Logistic}(\nu_{w,t} + \mu_{r,t} + \eta_{w,r,t}) \\
\eta_{w,r,t} &\sim \text{Normal}(\sum_{r'} a_{r' \to r} \eta_{w,r',t-1}, \gamma_{w,r})
\end{aligned}
$$

- $a_{i \to j}$ captures the linguistic "influence" of city i on city j.
- If $\eta_{j,t+1} = \eta_{i,t}$, then $a_{i \to j} = 1$, and $a_{i \to j} = 0$.
- If η_j and η_i co-evolve smoothly, then $a_{i,j} > 0$ and $a_{j,i} > 0$.

Language change as an autoregressive process



$\eta_2 \sim N(A\eta_1, \Sigma)$



$\eta_3 \sim N(A\eta_2, \Sigma)$

$$
\begin{aligned}
c_{\text{ctfu},1} &\sim \text{Binomial}(f(\eta_{\text{ctfu},1}), N_1) \\
c_{\text{hella},1} &\sim \text{Binomial}(f(\eta_{\text{hella},1}), N_1) \\
&\dots
\end{aligned}
$$

$$
\begin{aligned}
c_{\text{ctfu},2} &\sim \text{Binomial}(f(\eta_{\text{ctfu},2}), N_2) \\
c_{\text{hella},2} &\sim \text{Binomial}(f(\eta_{\text{hella},2}), N_2) \\
&\dots
\end{aligned}
$$

Estimating parameters of this autoregressive process reveals geographic pathways of diffusion across thousands of words.

Inference

$$P(\text{words}; \text{influence}) \triangleq P(c; a)$$

$$= \sum_z P(c, z; a) = \sum_z \overbrace{P(c \mid z)}^{\text{emission}} \overbrace{P(z; a)}^{\text{transition}}$$

(z represents "activation")

Inference

$$P(\text{words}; \text{influence}) \triangleq P(c; a)$$

$$= \sum_z P(c, z; a) = \sum_z \overbrace{P(c \mid z)}^{\text{emission}} \overbrace{P(z; a)}^{\text{transition}}$$

(z represents "activation")

$$= \int P(c \mid z) P(z; a) dz \qquad \text{(uh oh...)}$$

Inference

$$P(\text{words}; \text{influence}) \triangleq P(c; a)$$

$$= \sum_z P(c, z; a) = \sum_z \overbrace{P(c \mid z)}^{\text{emission}} \overbrace{P(z; a)}^{\text{transition}}$$

(z represents "activation")

$$= \int P(c \mid z) P(z; a) dz \qquad \text{(uh oh...)}$$

$$\rightarrow z^{(k)}, k \in \{1, 2, \ldots, K\}$$

$$\approx \sum_k P(c \mid z^{(k)}) P(z^{(k)}; a)$$

(Monte Carlo approximation to the rescue!)

Inference

$$P(\text{words; influence}) \triangleq P(c; a)$$

$$= \sum_z P(c, z; a) = \sum_z \overbrace{P(c \mid z)}^{\text{emission}} \overbrace{P(z; a)}^{\text{transition}}$$

(z represents "activation")

$$= \int P(c \mid z) P(z; a) dz \qquad \text{(uh oh...)}$$

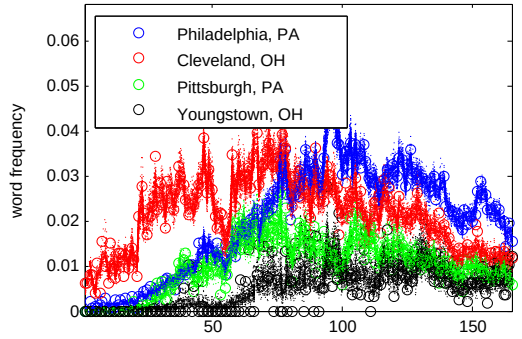
$$\rightarrow z^{(k)}, k \in \{1, 2, \ldots, K\}$$

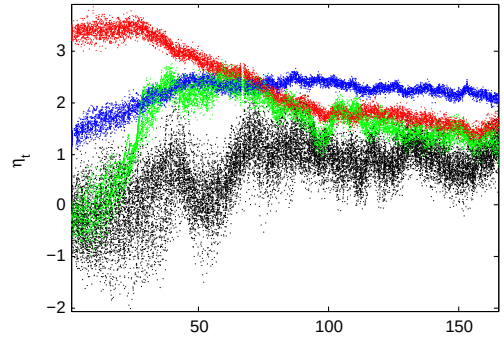
$$\approx \sum_k P(c \mid z^{(k)}) P(z^{(k)}; a)$$

(Monte Carlo approximation to the rescue!)

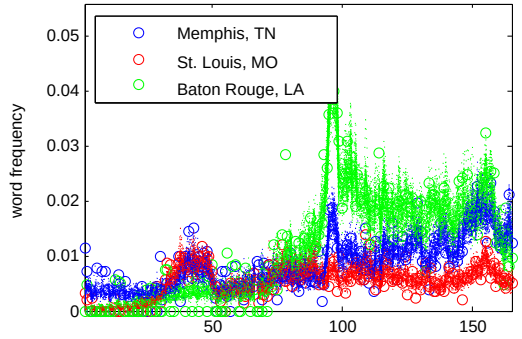
$$\hat{a} = \arg\max_a \sum_k P(c \mid z^{(k)}) P(z^{(k)}; a)$$

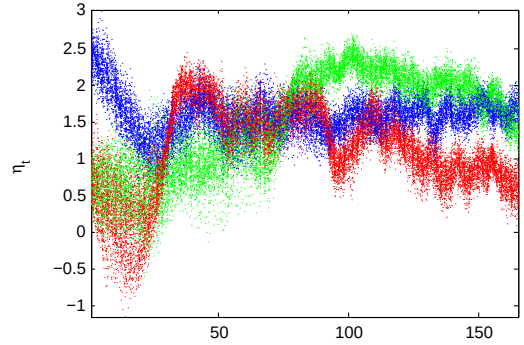
ctfu


Philadelphia, PA
Cleveland, OH
Pittsburgh, PA
Youngstown, OH
word frequency

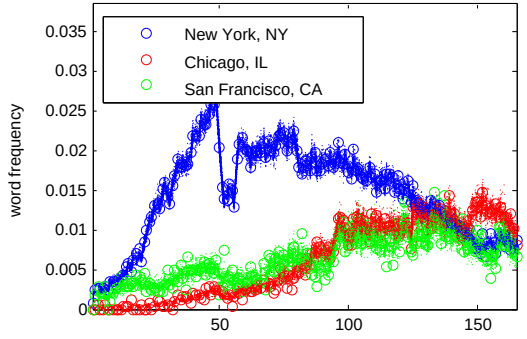

η_t

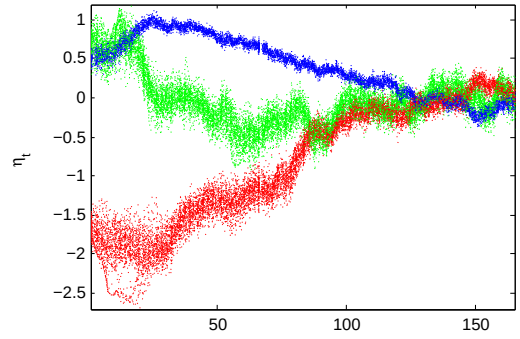
ion


Memphis, TN
St. Louis, MO
Baton Rouge, LA
word frequency

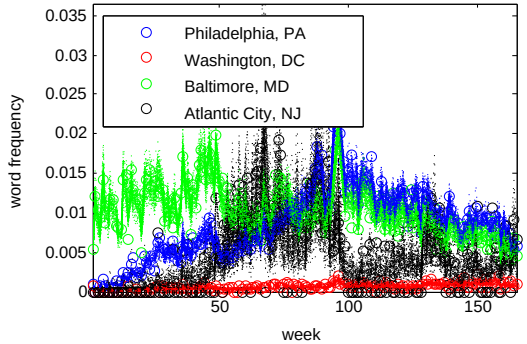

η_t

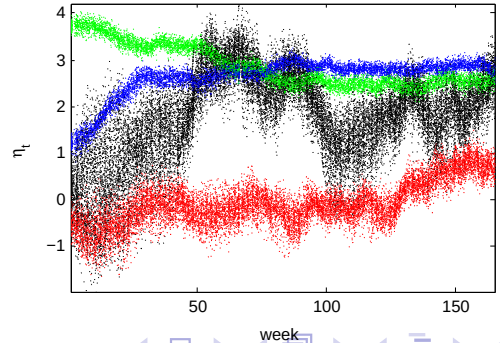
-__-


New York, NY
Chicago, IL
San Francisco, CA
word frequency

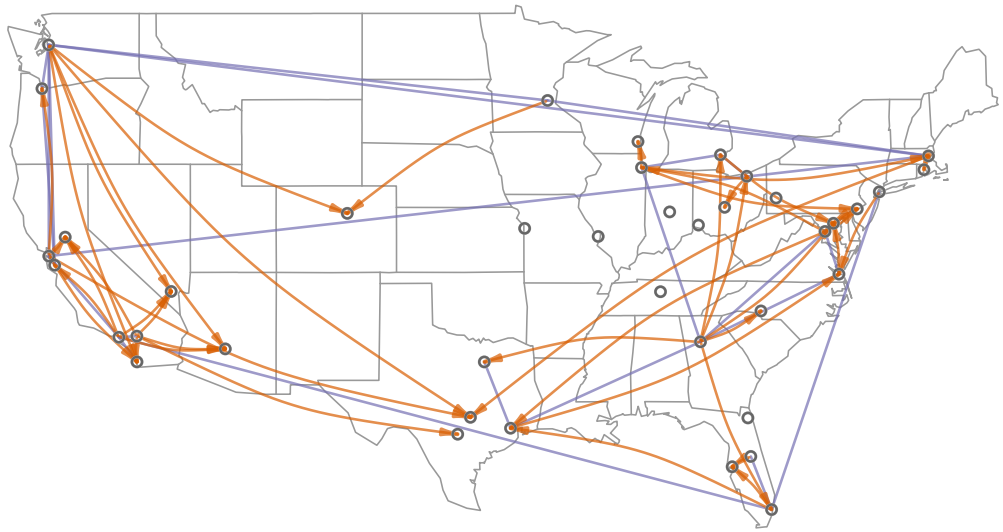

η_t

ard


Philadelphia, PA
Washington, DC
Baltimore, MD
Atlantic City, NJ
word frequency
week


η_t
week

Aggregating city-to-city autocorrelation



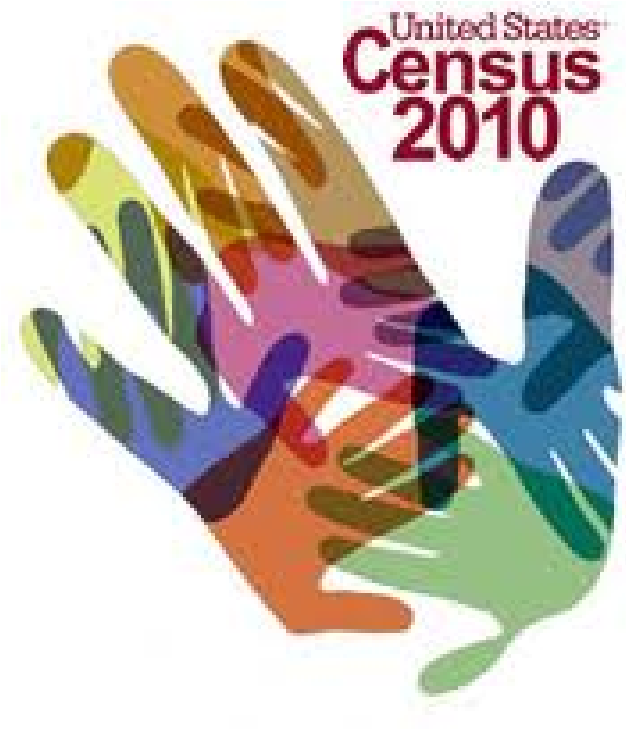
- Discretization by false discovery rate: 5% of links are false positives.
- But overall, linked cities are much closer than randomly-chosen cities.

Possible roles for demographics

- **Assortativity**: similar cities evolve together.
- **Influence**: certain types of cities tend to lead, others follow.

Possible roles for demographics

- **Assortativity**: similar cities evolve together.
- **Influence**: certain types of cities tend to lead, others follow.



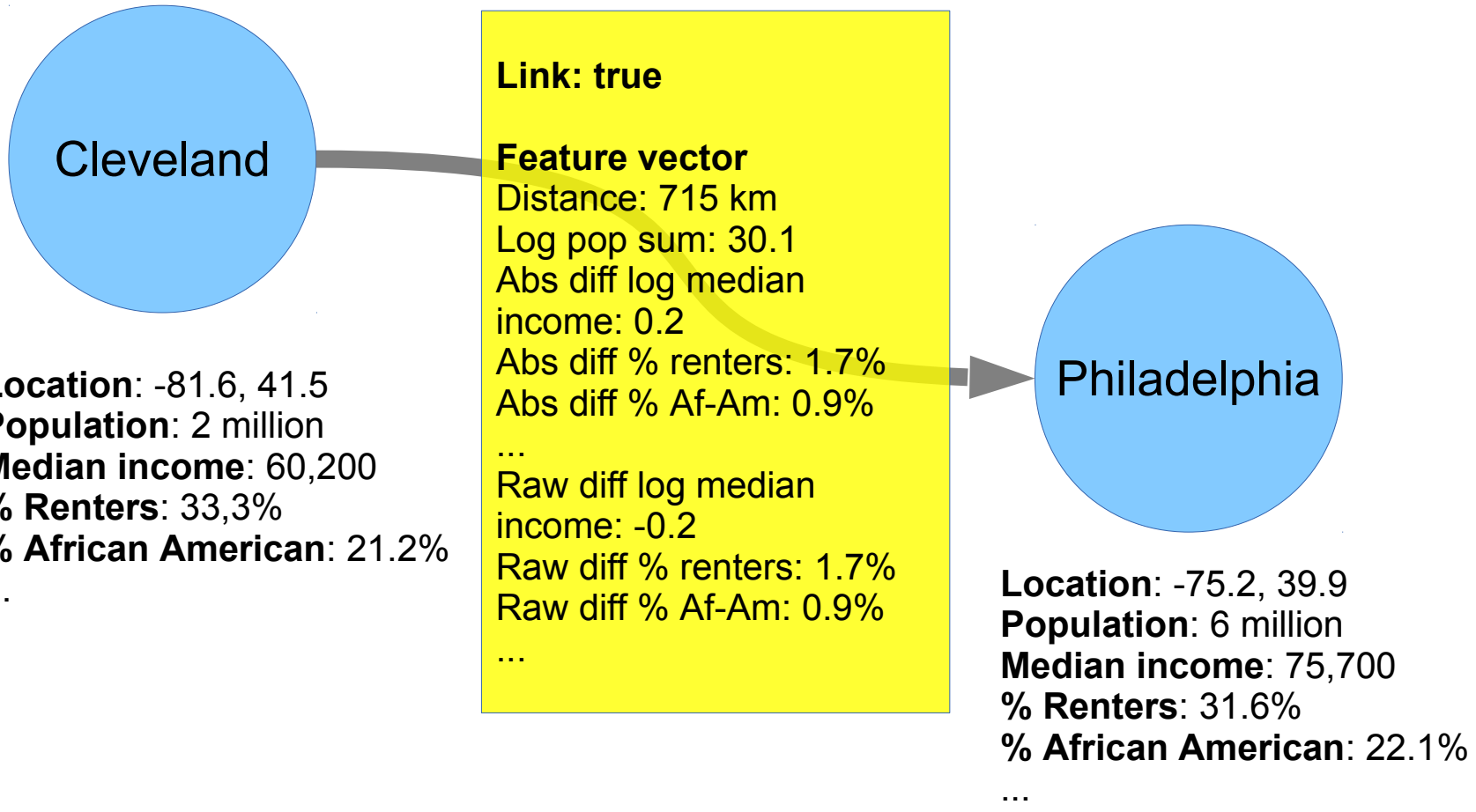
- 2010 US Census gives detailed demographics for each city.
- Are there types of demographic relationships that are especially frequent among linked cities?

Logistic regression

Location: -81.6, 41.5
Population: 2 million
Median income: 60,200
% Renters: 33,3%
% African American: 21.2%
...

Location: -75.2, 39.9
Population: 6 million
Median income: 75,700
% Renters: 31.6%
% African American: 22.1%
...

Logistic regression



Results

Coevolution

- Geography plays a strong role...
- ...but racial demographics is the trump card. Geographically odd couples:
  - DC $\leftrightarrow$ New Orleans
  - Los Angeles $\leftrightarrow$ Miami
  - Boston $\leftrightarrow$ Seattle

Influence: larger, younger cities tend to lead.

What would it take to get you to use a new word?

What would it take to get you to use a new word?

- Everyday linguistic decisions are based on social relationships and judgments.
- Aggregating millions of those decisions reveals the hidden structure of sociocultural influence.

What would it take to get you to use a new word?

- Everyday linguistic decisions are based on social relationships and judgments.
- Aggregating millions of those decisions reveals the hidden structure of sociocultural influence.

Thanks!

Brendan O'Connor

Noah Smith

Eric Xing

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2014). Diffusion of lexical change in social media. *PLOS-ONE*, *9*.