Language variation and change in social media

Jacob Eisenstein

Georgia Institute of Technology

April 13, 2013

Where I'm from

Natural language in computer science

twitter

- Natural language processing has focused on news text.
- Social media offers new opportunities, but poses serious linguistic challenges.

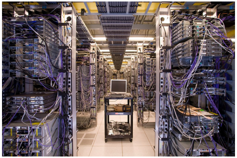
Some questions

What is the relationship between spoken language variation and language in social media?

- Do spoken language variables appear in social media?
- Does social media introduce new kinds of language variation?
- Does social media require reconsideration of social variables?
- How is social media language changing over time?

Why computers might help

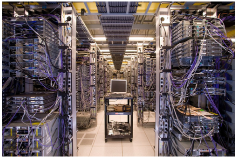
Social media corpora open the possibility of a new, “big data” methodology.



Why computers might help

Social media corpora open the possibility of a new, "big data" methodology.

- **Exploratory analysis**
  find linguistic variables in the data, rather than relying on experimenter's intuition.
- **Limited observer bias**
  language from real (public) social interactions, outside a lab.
- **Law of large numbers**
  a big pool of participants means less sensitivity to outliers.



What you might think Twitter looks like

What Twitter really looks like

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

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

Boom! Ya ur website suxx bro

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

What Twitter really looks like

It's not just celebrities:

- lol yeaa uu better! lol waht uu doin today?
- Love uu and miss you, sad I can't be there!
- Omqq =0 I Love uu Leel Wayne

What Twitter really looks like

It's not just celebrities:

- lol yeaa **uu** better! lol waht **uu** doin today?
- Love **uu** and miss you, sad I can't be there!
- Omqq =0 I Love **uu** Leel Wayne

uu is neither shorter nor easier to type than u.
Is it just a typo?

Here's looking at uu

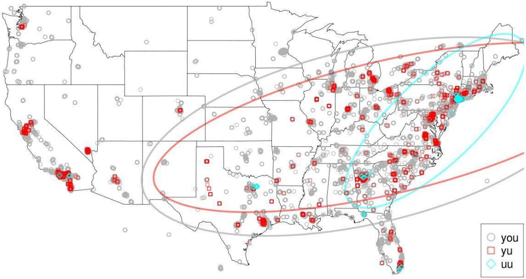
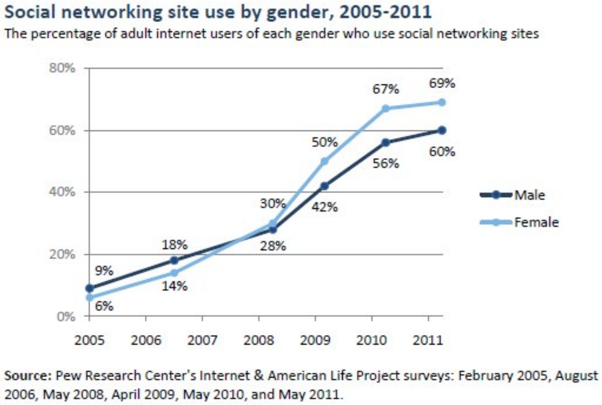


Figure: You and variants in March 2010

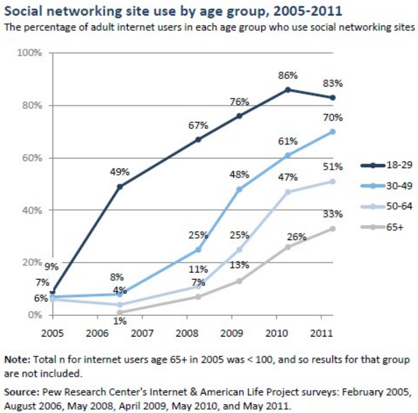
The spelling uu is strongly associated with New York, and rarely appears elsewhere (circa 2010).

Who uses social media?



(Pew Research Center, Aug 2011)

Who uses social media?

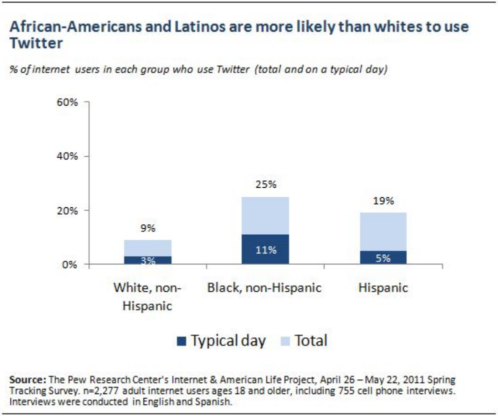


(Pew Research Center, Aug 2011)

Twitter

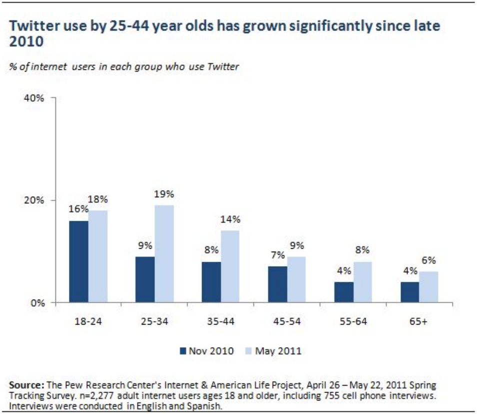
- Weak tie to real-life identity
- Short, unstructured content (140 characters)
- Unidirectional social network connections
- Can recover conversation traces from @-mentions

Who uses Twitter?



(Pew Research Center, June 2011)

Who uses Twitter?



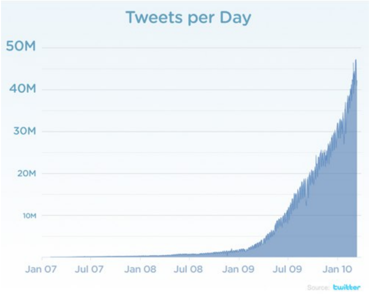
(Pew Research Center, June 2011)

Who uses Twitter on cellphones?

All cell owners (n=1954)	9%
Men (n=895)	9
Women (n=1059)	9
Age	
18-24 (n=225)	22**
25-34 (n=230)	14
35-44 (n=276)	9
45-54 (n=371)	5
55-64 (n=387)	3
65+ (n=429)	<1
Race/ethnicity	
White, Non-Hispanic (n=1404)	7
Black, Non-Hispanic (n=234)	17**
Hispanic (n=180)	12**

(Pew Research Center, May 2012)

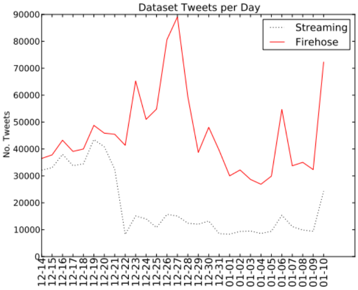
How much?



- Twitter claims:
  100 million active users, 177 million tweets per day in March 2011

Getting data from Twitter

- Twitter offers a sample of messages through their "streaming API."
- Supposedly 5% of all public messages, but not really.
- Unbiased sample? Nobody knows.
- Content cannot be redistributed.



(From Morstatter et al., ICWSM 2013)

Outline

Geographical Language Variation

A Latent Variable Model for Geographical Lexical Variation
Eisenstein, O'Connor, Smith, and Xing. EMNLP 2010.

- Does language display geographical variation in social media?
- If so, does it match spoken language variation?
- Where are the main linguistic divisions of the United States?
- Can their text predict where people are from?

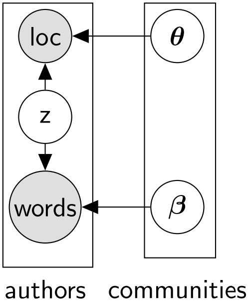
Dataset

- 9250 authors with GPS locations
- 380K messages from one week in March 2010
- 4.9M tokens
- Vocabulary limited to 5000 words (expanded later)
- Filters
  - At least 20 messages (in sample)
  - Must include GPS within a USA zipcode
  - No more than 1000 followers, followees

A mixture model for dialect

A very naïve model of where geotagged language comes from:

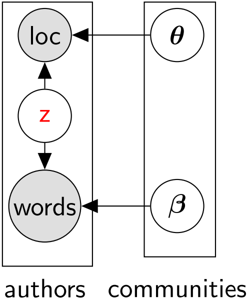
- Each author belongs to a geographical community.
- Each geographical community has probability distributions over words and locations.
- Each location is a random draw from a probability distribution associated with the author's community.
- Each author's text is a random draw from a probability distribution associated with the author's community.



A mixture model for dialect

A very naïve model of where geotagged language comes from:

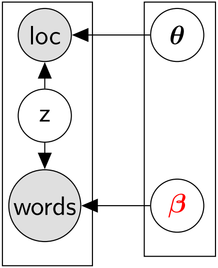
- Each author belongs to a geographical community.
- Each geographical community has probability distributions over words and locations.
- Each location is a random draw from a probability distribution associated with the author's community.
- Each author's text is a random draw from a probability distribution associated with the author's community.



A mixture model for dialect

A very naïve model of where geotagged language comes from:

- Each author belongs to a geographical community.
- Each geographical community has probability distributions over words and locations.
- Each location is a random draw from a probability distribution associated with the author's community.
- Each author's text is a random draw from a probability distribution associated with the author's community.

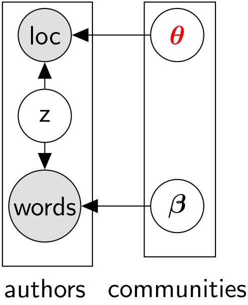


authors communities

A mixture model for dialect

A very naïve model of where geotagged language comes from:

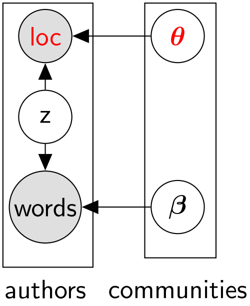
- Each author belongs to a geographical community.
- Each geographical community has probability distributions over words and locations.
- Each location is a random draw from a probability distribution associated with the author's community.
- Each author's text is a random draw from a probability distribution associated with the author's community.



A mixture model for dialect

A very naïve model of where geotagged language comes from:

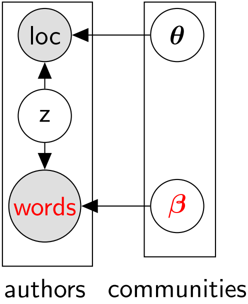
- Each author belongs to a geographical community.
- Each geographical community has probability distributions over words and locations.
- Each location is a random draw from a probability distribution associated with the author's community.
- Each author's text is a random draw from a probability distribution associated with the author's community.



A mixture model for dialect

A very naïve model of where geotagged language comes from:

- Each author belongs to a geographical community.
- Each geographical community has probability distributions over words and locations.
- Each location is a random draw from a probability distribution associated with the author's community.
- Each author's text is a random draw from a probability distribution associated with the author's community.



Limitations and extensions

- The mixture model assumes all lexical differences are either geographical, or IID noise.
- To account for non-geographical variation, add **latent topics**: groups of words which are used by the same authors.

Limitations and extensions

- The mixture model assumes all lexical differences are either geographical, or IID noise.
- To account for non-geographical variation, add **latent topics**: groups of words which are used by the same authors.
  - album, music, beats, artist, video, #lakers, itunes, tour
  - bieber, justin, gaga, jonas, pants, beiber, ring, annoying
  - da, dat, dis, wat, dats, dey, gud, watz, wats

Predictive accuracy

We can compute the **expected** location of each author by taking a weighted sum over geographical communities.

error in kilometers $\rightarrow$	**mean**	**median**
population center	1148	1018

Predictive accuracy

We can compute the **expected** location of each author by taking a weighted sum over geographical communities.

error in kilometers $\rightarrow$	**mean**	**median**
population center	1148	1018
text regression	948	712
supervised LDA	1055	728

Predictive accuracy

We can compute the **expected** location of each author by taking a weighted sum over geographical communities.

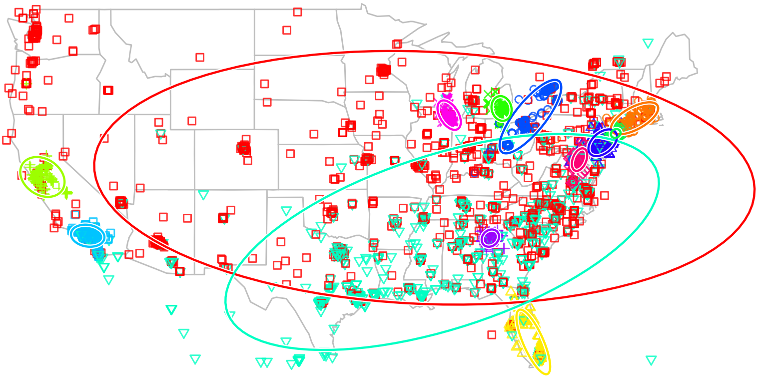
error in kilometers $\rightarrow$	**mean**	**median**
population center	1148	1018
text regression	948	712
supervised LDA	1055	728
mixture model	947	644
+topics	900	494

Predictive accuracy

We can compute the **expected** location of each author by taking a weighted sum over geographical communities.

error in kilometers $\rightarrow$	**mean**	**median**
population center	1148	1018
text regression	948	712
supervised LDA	1055	728
mixture model	947	644
+topics	900	494
+sparsity [EAX'11]	845	501
+larger vocab	791	461

Text+geography model output



Text+geography model output

For each cluster,[1] rank words by log-odds: $\log \boldsymbol{\beta}_i - \log \frac{1}{K} \sum_j \boldsymbol{\beta}_j$:

- **New York**: brib, lml, wassupp, uu, werd, deadass, flatbush, odee, dha
- **So. Cal**: disneyland, cuh, fucken, af, fasho, faded, wyd, freeway, bomb
- **No. Cal**: sac, oakland, sf, hella, warriors, pleasure, bay, koo
- **Atlanta**: atlanta, atl, georgia, ga, $1, waffle, af, nun, shawty
- **Cleveland/Detroit**: ctfu, detroit, foolin, .!!, cleveland, geeked, salty, ikr
- **Northwest**: seattle, portland, oregon, olympic, heh, canada, stoked

[1] note: clusters do not match previous slide.

Text+geography model output

For each cluster,[1] rank words by log-odds: $\log \boldsymbol{\beta}_i - \log \frac{1}{K} \sum_j \boldsymbol{\beta}_j$:

- **New York**: brib, lml, wassupp, uu, werd, deadass, flatbush, odee, dha
- **So. Cal**: disneyland, cuh, fucken, af, fasho, faded, wyd, freeway, bomb
- **No. Cal**: sac, oakland, sf, hella, warriors, pleasure, bay, koo
- **Atlanta**: atlanta, atl, georgia, ga, $1, waffle, af, nun, shawty
- **Cleveland/Detroit**: ctfu, detroit, foolin, .!!, cleveland, geeked, salty, ikr
- **Northwest**: seattle, portland, oregon, olympic, heh, canada, stoked

[1] note: clusters do not match previous slide.

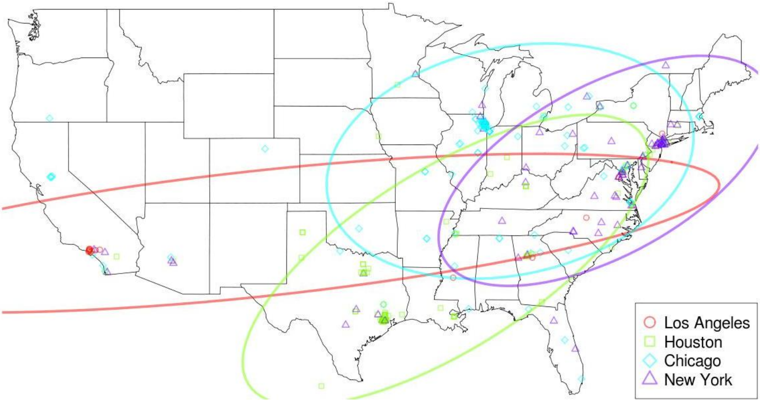
Text+geography model output

For each cluster,[1] rank words by log-odds: $\log \boldsymbol{\beta}_i - \log \frac{1}{K} \sum_j \boldsymbol{\beta}_j$:

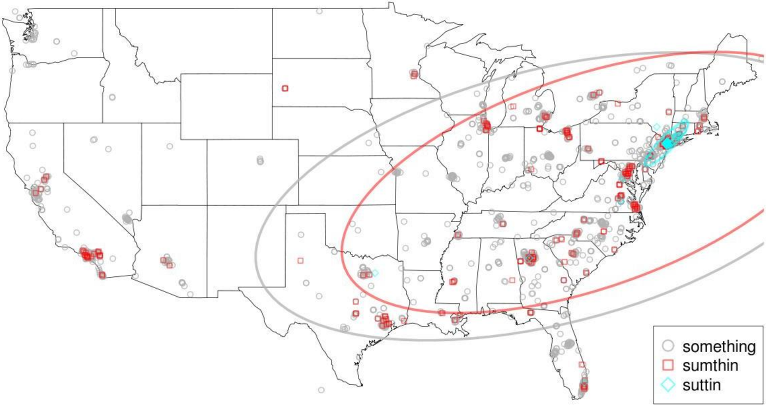
- **New York**: brib, lml, wassupp, uu, werd, deadass, flatbush, odee, dha
- **So. Cal**: disneyland, cuh, fucken, af, fasho, faded, wyd, freeway, bomb
- **No. Cal**: sac, oakland, sf, hella, warriors, pleasure, bay, koo
- **Atlanta**: atlanta, atl, georgia, ga, $1, waffle, af, nun, shawty
- **Cleveland/Detroit**: ctfu, detroit, foolin, .!!, cleveland, geeked, salty, ikr
- **Northwest**: seattle, portland, oregon, olympic, heh, canada, stoked

[1] note: clusters do not match previous slide.

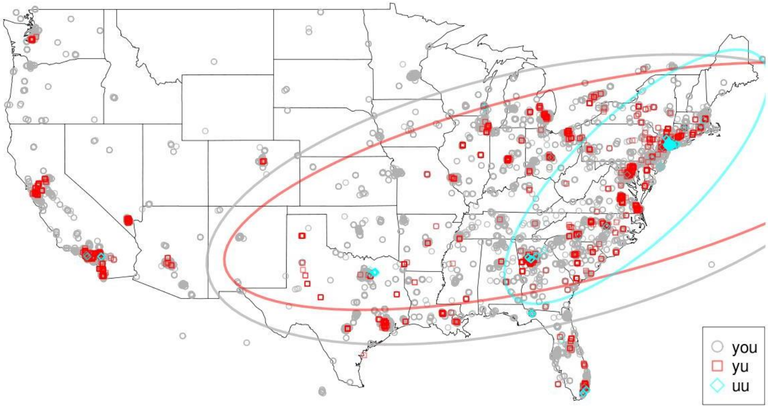
Mentions of city names



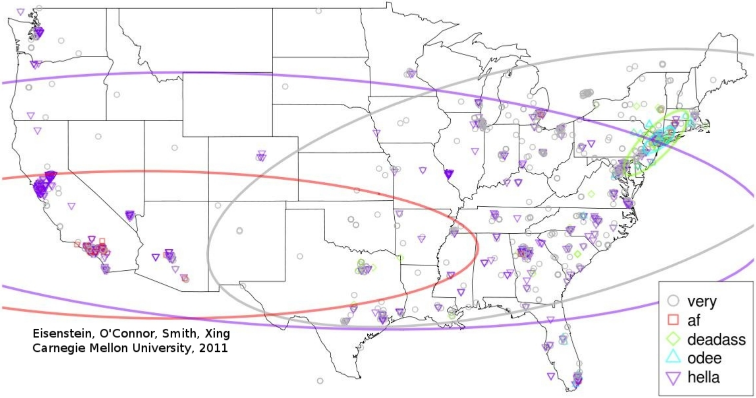
Something and variants



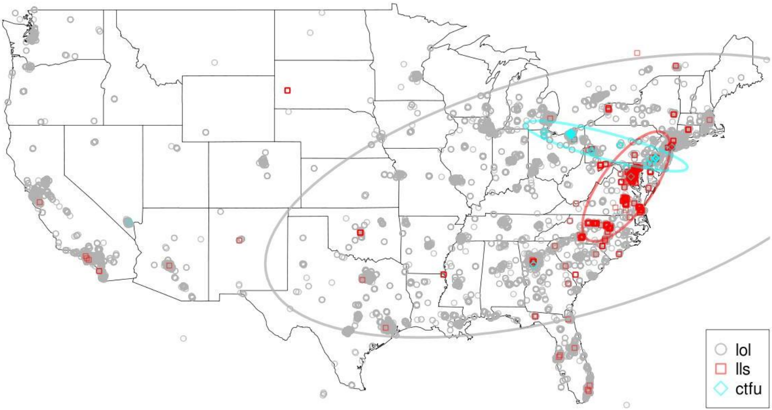
You and variants



Intensifiers



LOL and variants



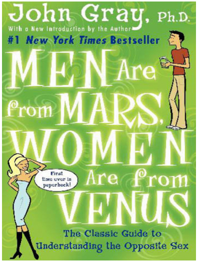
Outline

Predictive models and personal identity

- According to our analysis, Southern California is characterized by words like af, fasho, bomb.
- My sister-in-law is a Southern Californian lifer.
  She never uses these words!
- So you have an accurate predictive model...
  What kind of descriptive statements does that license?

Language and gender

- Lots of research in **predicting** gender from social media text.
- Predictions are $\sim 90\%$ accurate.
- The descriptive analysis is... uninspiring (unless you love traditional gender roles).
  - Men prefer "content," women prefer "style." (Argamon et al. 2003, 2007)
  - Women prefer "expressive" words. (Rao et al. 2010, Burger et al. 2011)



Alternative identities

Gender identity and lexical variation in Twitter
Bamman, Eisenstein, and Schnoebelen. In preparation.

- We started with a painfully simple idea:
  - Language use reflects gender

    88% accuracy from bag-of-words
  - Social networks are often homophilous with respect to gender.
  - Can we put these two features together to accurately predict the gender of authors on Twitter?

Data

- 14,464 Twitter users (56% male)
  - geolocation in USA
  - must use 50 of 1000 most frequent words
  - no more than 1000 follow connections
- 9.2M tweets, from January to June 2011
- Author gender induced from given name and census records.

  The median author's name is 99.6% homogeneous
- Social network induced from mutual @-mentions
  - Women have 58% female friends
  - Men have 67% male friends

Why does classification work?

	F	M
Standard dictionary	74.2%	74.9%
Punctuation	14.6%	14.2%
Non-standard, unpronounceable words (e.g., :), lmao)	4.28%	2.99%
Non-standard, pronounceable words (e.g., luv)	3.55%	3.35%
Named entities	1.94%	2.51%
Numbers	0.83%	0.99%
Taboo	0.47%	0.69%
Hashtags	0.16%	0.18%

Table: Word category frequency by gender. All differences are statistically significant at $p < .01$.

Clustering by content

- At the corpus level, women use more non-dictionary words and men mention more named entities.
- But are "men" and "women" the right categories?
- We performed a clustering over all authors by text.
  - K-means ($K = 20$)
  - Clusters represent shared interests and/or styles.

    It's not so easy to pull these apart...
  - Many clusters **happen to have** strong demographic orientations, including gender.

Female clusters

% fem	words
0.84	fabric blogged hubs recipe recipes delish @starbucks almond howdy baking cocktails
0.79	;o xx hun xxx hump sweetie x xoxoxo cena becky
0.78	xo elizabeth gr8 -) ranked ty blessings thnx fr 2day
0.76	muah darren bo sry xoxoxo sux ,,, scotty lmbo hun
0.75	clark pokemon ash arc #idol authors unicorns terrifying romance chapter
0.75	:') (: <333 @justinbieber (; xxx <33333 </3 <33 ;d

Female clusters

% fem	words
0.84	fabric blogged hubs recipe recipes delish @starbucks almond howdy baking cocktails
0.79	;o xx hun xxx hump sweetie x xoxoxo cena becky
0.78	xo elizabeth gr8 -) ranked ty blessings thnx fr 2day
0.76	muah darren bo sry xoxoxo sux ,,, scotty lmbo hun
0.75	clark pokemon ash arc #idol authors unicorns terrifying romance chapter
0.75	:') (: <333 @justinbieber (; xxx <33333 </3 <33 ;d

- At the population level, women use few named entities and many non-dictionary words.
- But there are clusters of (mostly) women who do the opposite.

Male clusters

% fem	words
0.29	dems gop democrats unions conservative senate muslim israel liberal republicans
0.28	niggaz shyt dats dey wats lmmfao lik dis neva lls
0.19	e3 gears psn 360 kombat halo gaming portal console marvel
0.19	bama @darrenrovell @espn severe auburn ky #heat thunderstorm au #marchmadness
0.15	#nba mets #jets #mavs #knicks crawford @ochocinco pacers #lakers wright
0.14	api ui ios apple's developers developer dev hardware plugin interface
0.07	#nhl nhl prospect #bruins qb roster timeout 2-1 boozer 1-0

- At the population level, men use many named entities and few non-dictionary words.
- But there are clusters of men who do the opposite.

What about the people that we got wrong?

- 88% accuracy means 12% errors.
- Can we fix those errors by adding new information?
- Social network homophily:
  63% of @-mentions are between same-gender individuals.
- Maybe social network features will disambiguate errors made by the language features.

Adding social network features

Logistic regression, 10-fold cross-validation:

- Text alone: 88% accuracy

Adding social network features

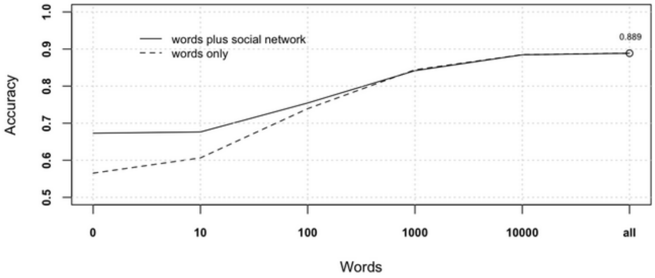
Logistic regression, 10-fold cross-validation:

- Text alone: 88% accuracy
- Text+network: 88% accurate

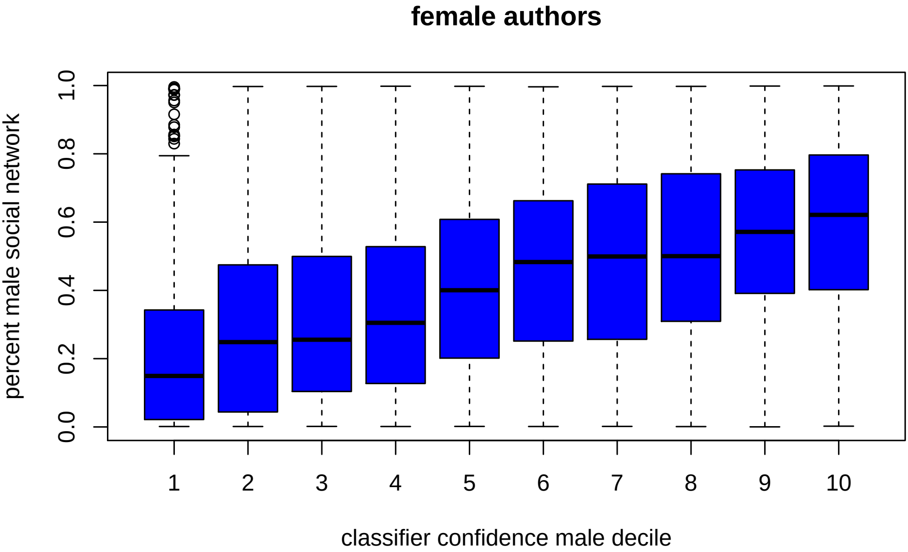
Adding social network features

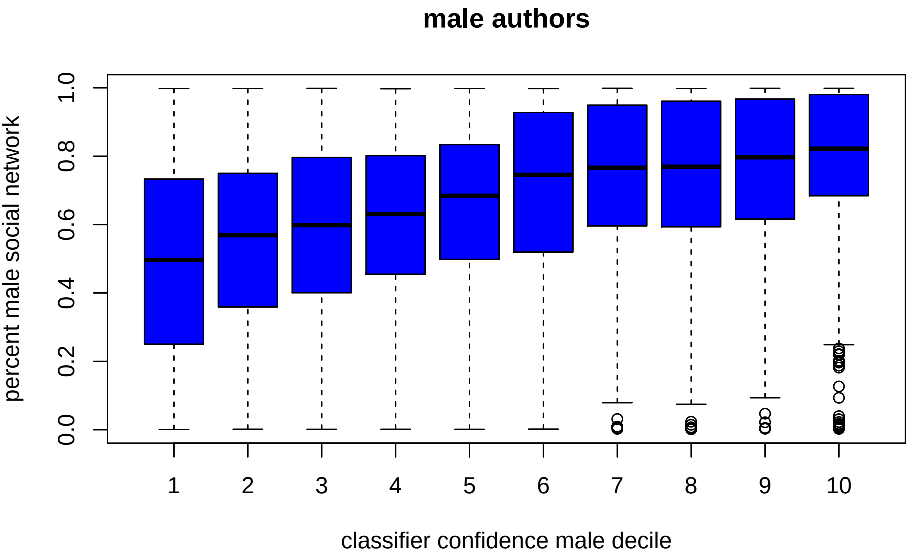
Logistic regression, 10-fold cross-validation:

- Text alone: 88% accuracy
- Text+network: 88% accurate



Once we have 1000 words per author, adding network information does not improve performance. **Why not?**


female authors
percent male social network
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
7
8
9
10
classifier confidence male decile


male authors
percent male social network
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
7
8
9
10
classifier confidence male decile

Why social network features don't help

correlation	female authors	male authors
classifier vs. network	0.38 ($.35 \leq r \leq .40$)	0.33 ($.30 \leq r \leq .36$)

Why social network features don't help

correlation	female authors	male authors
classifier vs. network	0.38 ($.35 \leq r \leq .40$)	0.33 ($.30 \leq r \leq .36$)

- Network features will improve gender classification only to the extent that they are adding new information.

Why social network features don't help

correlation	**female authors**	**male authors**
classifier vs. network	0.38 ($.35 \leq r \leq .40$)	0.33 ($.30 \leq r \leq .36$)

- Network features will improve gender classification only to the extent that they are adding new information.
- But language and social network are correlated even after controlling for author gender.

Summary

Cluster analysis: There are broad language differences between genders, but large clusters individuals “violate” overall norms.

Summary

Cluster analysis: There are broad language differences between genders, but large clusters individuals "violate" overall norms.

- Writing like a woman or a man doesn't mean one thing: gender interacts with other social variables in complex ways.
- Accurate prediction of a social attribute does not license blanket statements about its linguistic characteristics.

Summary

Cluster analysis: There are broad language differences between genders, but large clusters individuals "violate" overall norms.

- Writing like a woman or a man doesn't mean one thing: gender interacts with other social variables in complex ways.
- Accurate prediction of a social attribute does not license blanket statements about its linguistic characteristics.

Social network analysis: Linguistic and social network gender predictors are correlated, *even when holding gender constant*.

- Rather than seeing these features as revealing the author's "true" gender, they reveal an attitude towards gender.

Outline

Social media variation and spoken language variation

Phonological factors in social media writing. Eisenstein 2013.

- What is the relationship between spoken language variation and social media writing?
  - Some replication of known lexical variables

    hella, jawn
  - Some variables seem specific to written language

    ctfu, uu
  - Some seem to have something to do with spoken language...

    suttin (something), shawty (shorty), wassup (what's up)
- Does spoken language variation interact with social media writing in a systematic way?

Final consonant deletion in Twitter

left / lef	ok **lef** the y had a good workout
just / jus	**jus** livin this thing called life

Final consonant deletion in Twitter

left / lef	ok **lef** the y had a good workout
just / jus	**jus** livin this thing called life
with / wit	da hell **wit** u

Final consonant deletion in Twitter

left / lef	ok **lef** the y had a good workout
just / jus	**jus** livin this thing called life
with / wit	da hell **wit** u
going / goin	when is she **goin** bck 2 work?
doing / doin	he **doin** big things

Final consonant deletion in Twitter

left / lef	ok **lef** the y had a good workout
just / jus	**jus** livin this thing called life
with / wit	da hell **wit** u
going / goin	when is she **goin** bck 2 work?
doing / doin	he **doin** big things
know / kno	u **kno** u gotta put up pics

African American English in writing

- (TD)-deletion is associated with several regional and ethnic dialects, particularly AAE (Labov 1968, Green 2002)
- Earlier studies found little evidence of phonological features of AAE in writing:
  - Whiteman (1982)

    > *Nonstandard phonological features [of AAE] rarely occur in writing, even when those features are extremely frequent in the oral dialect of the writer.*

  - Thompson et al (2004)

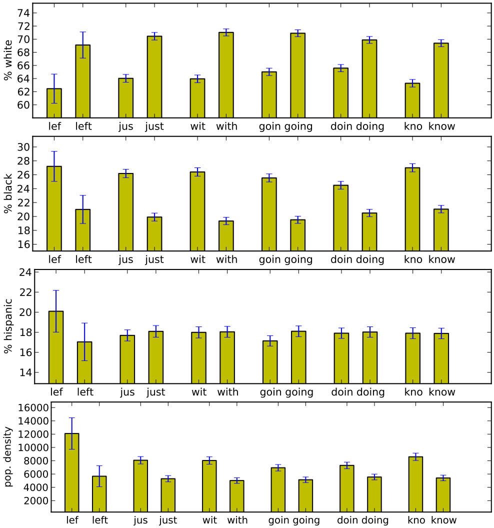
    > *African American students have models for* **spoken** *AAE; however, children do not have models for written AAE... students likely have minimal opportunities to experience AAE in print.*

Who is dropping final consonants?

Aggregate census statistics as a proxy for author demographics:

- Find average geographic coordinates for each author.
- Identify five-digit US census block.
- Compute average demographic profile.

The demographics of final consonant deletion



- (TD)-deletion occurs in census blocks with:
    - more African Americans,
    - fewer European Americans,
    - and greater population density...
- But so does every other kind of final consonant deletion!

When are they dropping final consonants?

- In speech, (TD)-deletion is inhibited when preceding vowel-initial segments (e.g., Guy 1991).
  - She **lef** the keys
  - She **left** a tip
- Does consonant dropping in Twitter also depend on context?
- Raw frequencies are confounded by a few very frequent expressions, e.g. going to, mos def
- Logistic regression
  - Dependent variable: final consonant deletion
  - Independent variable: does next segment start with a vowel?
  - "Random effects" for each subsequent word

Logistic regression

	μ_β	σ_β	z	p
lef / left	-0.45	0.10	-4.47	3.9×10^{-6}
jus / just	-0.43	0.11	-3.98	3.4×10^{-5}
wit / with	-0.16	0.03	-4.96	3.6×10^{-7}
doin / doing	0.08	0.04	2.29	0.011
goin / going	-0.07	0.05	-1.62	0.053
kno / know	-0.07	0.05	-1.23	0.11

Table: Logistic regression coefficients for the VOWEL feature, predicting the choice of the shortened form.

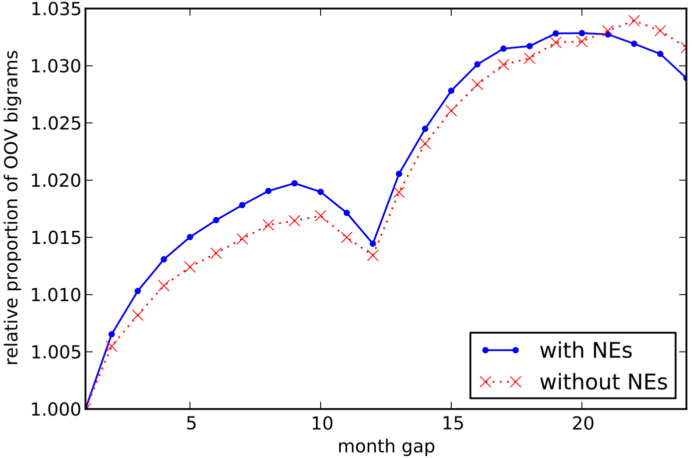
Contextual influences on consonant deletion

A role for phonological factors in social media writing?

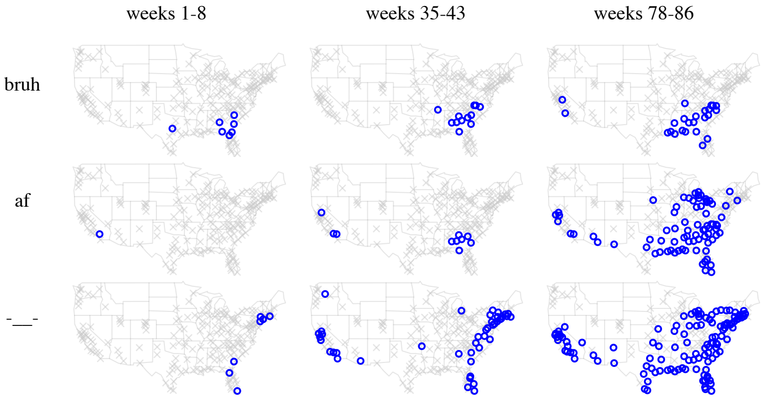
- The consonsanat deletions in lef, jus, and wit are significantly **less** likely when followed by a vowel.
- doin is **more** likely when followed by a vowel.
- These contextual factors are evidence against purely lexical account of variation in social media text.

Outline

Language in social media is constantly changing



New words over time and space



Blue circles are cities in which the word is used by at least 1% of the people who post to Twitter in a given week.

Modeling the spread of new words

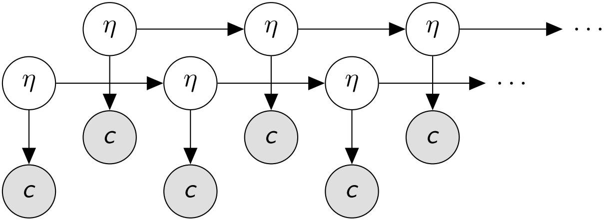
Mapping the geographical diffusion of new words. Eisenstein, O'Connor, Smith, Xing. In preparation.

- Measure word frequency in 200 American cities over two years.
- Aggregate across thousands of words to obtain a single model of city-to-city linguistic influence.
- A large-scale real-time empirical testbed for theories of language change.
  - 44 million messages. Mostly English; no retweets; no URLs.
  - 495,000 authors, all geolocated to an American city (MSA)
  - Two years of text, coarsened to one-week bins

Language change as a linear dynamical system

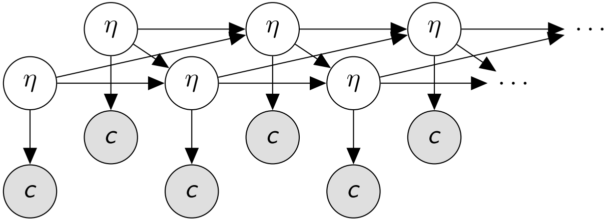
- Power law distributions over both word frequency and city size
- Counts of rare words in small cities will be sparse, making estimation challenging.
- We propose a linear dynamical system, treating the popularity of a word in a city as a latent variable.

Language change as a linear dynamical system



- Word counts c_{rti} are drawn from a Binomial distribution, whose parameter incorporates:
  - Overall popularity of word i at time t
  - Overall verbosity of region r at time t
  - "Extra" word-specific popularity, $\eta_{i,r,t}$
- Latent popularity evolves as $\eta_{i,r,t} = A\eta_{i,r,t-1} + \epsilon_{i,r,t}$.

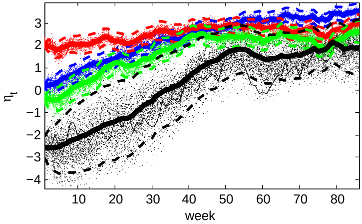
Language change as a linear dynamical system

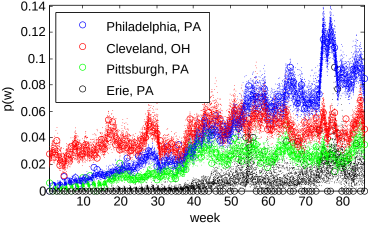


- Word counts c_{rti} are drawn from a Binomial distribution, whose parameter incorporates:
  - Overall popularity of word i at time t
  - Overall verbosity of region r at time t
  - "Extra" word-specific popularity, $\eta_{i,r,t}$
- Latent popularity evolves as $\eta_{i,r,t} = A\eta_{i,r,t-1} + \epsilon_{i,r,t}$.
- Off-diagonal elements in A represent cross-regional influence.

Managing uncertainty

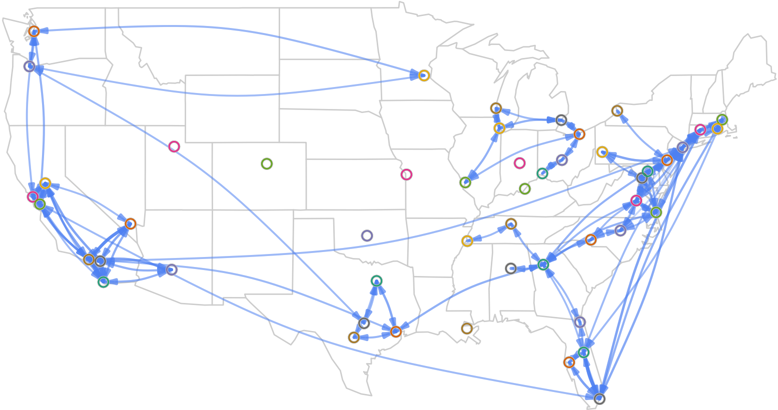
- When word counts are large, we trust our estimates of A.
- But counts of rare words in small cities will be sparse, due to power law distributions.
- We use sequential Monte Carlo to approximate $P(\eta|c)$ with a set of samples.





- We can estimate the influence matrix A in each sample, and fit a Gaussian to the set of estimates.

Network of influence



Geography plays no explicit role in constructing the network, but most influence links are between geographically proximate cities.

Reading tea leaves?

What types of cities share influence?

Logistic regression to distinguish linked versus non-linked city pairs:

	β	t
product of pops	0.138	3.615
geo distance	-1.542	-18.126
difference features		
pct urbanized	-0.355	-6.966
median income	0.004	0.074
median age	-0.109	-1.904
% renter	-0.013	-0.252
% af. am	-0.866	-13.256
% hispanic	-0.013	-0.201

What factors make cities lead or follow?

Logistic regression to predict the leader in an asymmetric pair:

	β	t
log pop diff	1.03	6.48
pct urbanized	4.2e-3	0.338
median income	2.3e-5	2.498
median age	-5.4e-2	-1.217
% renter	-1.89e-2	-0.877
% af. am	3.67e-2	2.486
% hispanic	1.78e-2	1.791

In asymmetric relationships, the city that leads is usually larger, wealthier, and has more African Americans.