

Is Netspeak a Dialect?

Jacob Eisenstein
@jacobeisenstein

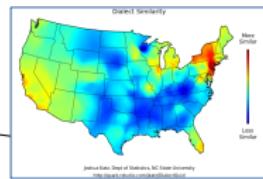
Georgia Institute of Technology

February 15, 2015

Variation by speaker and by medium

Dialect variation

geography, ethnicity, class, gender, ...



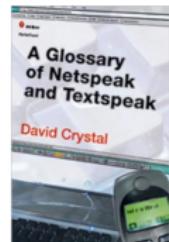
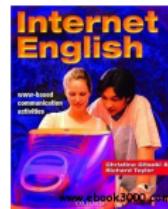
Variation by speaker and by medium



Standard writing

Variation by Medium

Netspeak

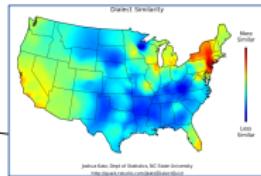


Variation by speaker and by medium



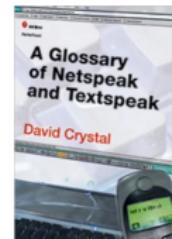
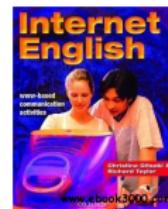
Standard writing

Variation by Medium



Dialect variation
geography, ethnicity, class, gender, ...

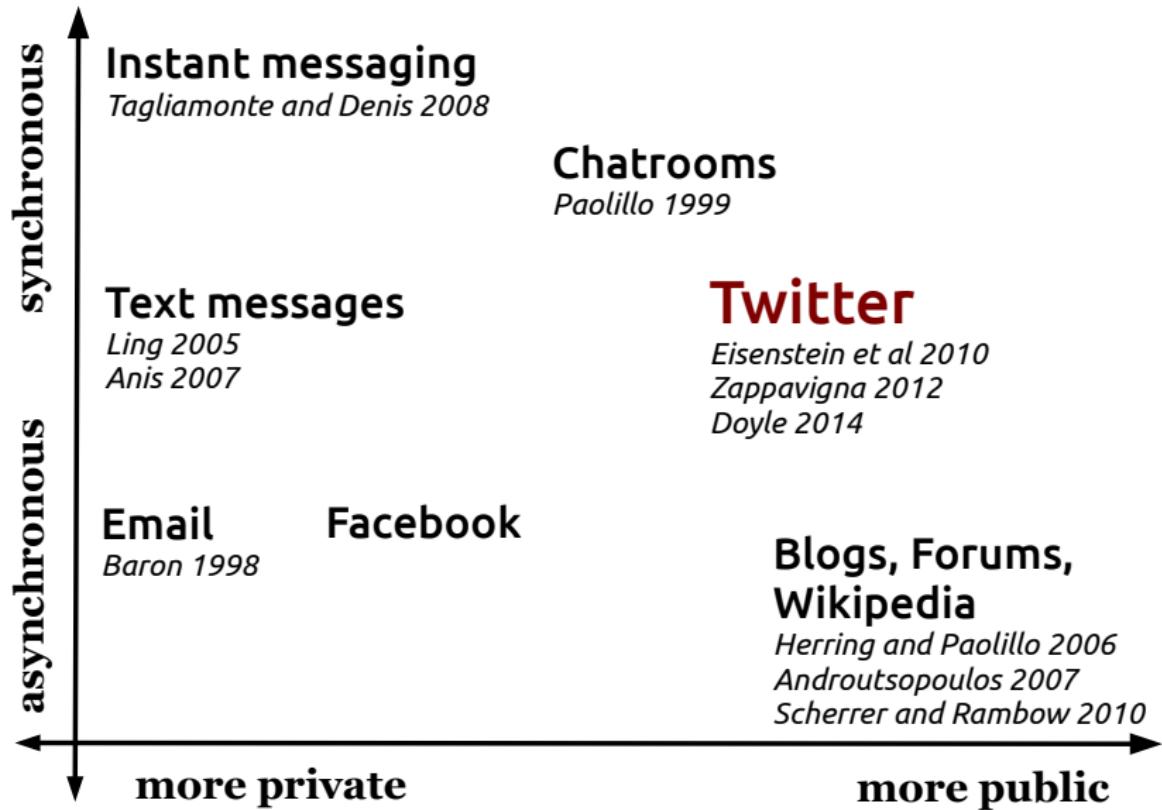
Netspeak



Questions for this talk

- ▶ Are spoken dialect features transcribed in digitally-mediated writing?
- ▶ How can we use large digital corpora to automatically induce dialect difference?
- ▶ Why would “netspeak” features vary with geography?

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, and social network and geographical metadata is available.

A screenshot of a Twitter feed showing four tweets from different users:

- AXIOM SFD** @AXIOMSFD - Aug 7
Improve your sales team performance by bringing together #CRM, learning & development, & **big data** insights bit.ly/lmq5n4
- Susan Visser** @susvis - Aug 7
Webinar: How to Mitigate Fraud and Cyber Threats with **Big Data** and Analytics: [#FraudPrevention #fintech](http://bit.ly/bdatafraud)
- giulio quaggiotto** @gquaggiotto - Aug 7
What uses for #**bigdata** in #globaldev? Getting ready for tomorrow's webinar with @ADB_HQ colleagues
- Communitelligence** @CommIntelligence - Aug 6
Measurement in the Age of Mobile, Sensors, **Big Data** and Google Glass webinar by Katie Paine ow.ly/A0XBm

A screenshot of a Twitter interface showing a sidebar titled "For Special Consideration: Twitter".

The sidebar contains the following text:

For Special Consideration: Twitter.

hahaha @ha_ha ha ha! #hahaha

hahahaha RT @ha_ha @hahaha ha ha! #hahaha, #hahahaha

hahhah RT @ha_ha @hahaha @hahahaha ha ha! #hahaha, #hahahaha, #hee_hee

yello_kOtAkU RT @ha_ha @hahaha @hahaha @hahahaha @hahahaha ha ha! #hahaha (trending), #hahahaha, #hee_hee, #wahaha

Questions for this talk

- ▶ Are spoken dialect features transcribed in digitally-mediated writing?
- ▶ How can we use large digital corpora to automatically induce dialect difference?
- ▶ Why would “netspeak” features vary with geography?

Questions for this talk

- ▶ **Are spoken dialect features transcribed in digitally-mediated writing?**
- ▶ How can we use large digital corpora to automatically induce dialect difference?
- ▶ Why would “netspeak” features vary with geography?

Variation in digital writing

Linguistic variables

- ▶ lexical items from speech

Social variables

- ▶ geography

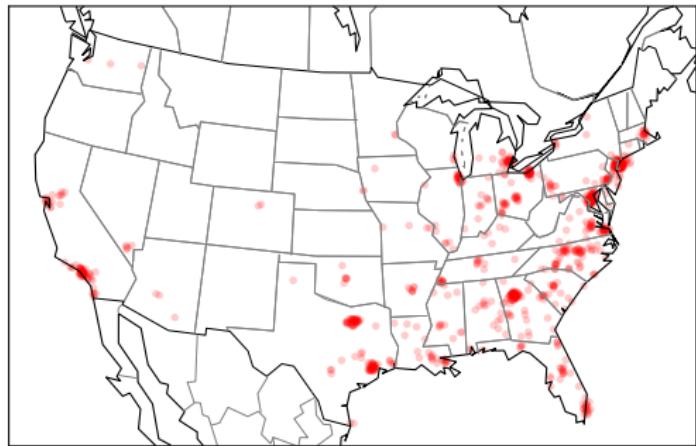
Yinz

- ▶ 2nd-person pronoun
- ▶ Western Pennsylvania
- ▶ Very rare: appears in 535 of 10^8 messages



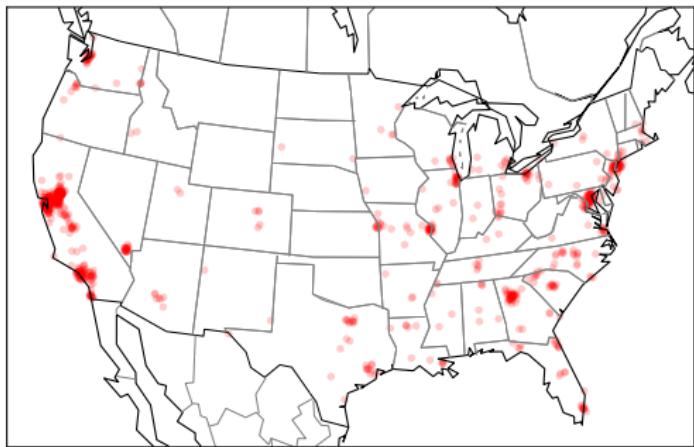
Yall

- ▶ 2nd-person pronoun
- ▶ Southeast, African-American English
- ▶ Once per 250 messages



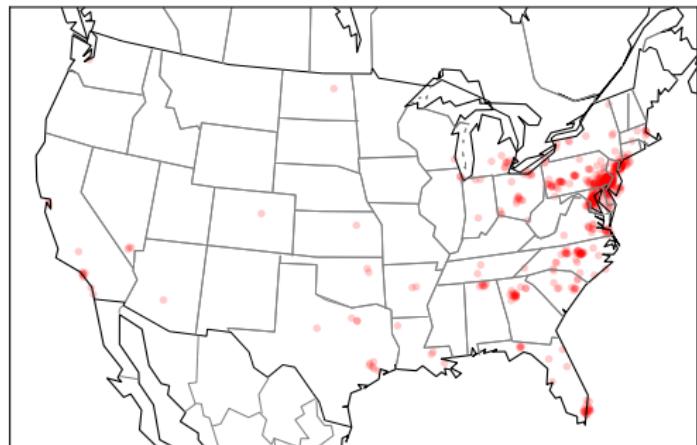
Hella

- ▶ Intensifier, e.g.
i got hella nervous
- ▶ Northern California¹
- ▶ Once per 1000 messages



Jawn

- ▶ Noun, diffuse semantics
- ▶ Philadelphia, hiphop²
- ▶ Once per 1000 messages



- ▶ @user ok u have heard this jawn right
- ▶ i did wear that jawn but it was kinda warm this week

Summary of spoken dialect terms

	rate	region
yinz	200,000	mainly used in Western PA
yall	250	ubiquitous
hella	1000	ubiquitous, but more frequent in Northern California
jawn	1000	mainly used in Philadelphia

- ▶ Overall: mixed evidence for spoken language dialect variation in Twitter.
- ▶ But are these the right words?

Questions for this talk

- ▶ **Are spoken dialect features transcribed in digitally-mediated writing?**
- ▶ How can we use large digital corpora to automatically induce dialect difference?
- ▶ Why would “netspeak” features vary with geography?

Questions for this talk

- ▶ Are spoken dialect features transcribed in digitally-mediated writing?
- ▶ **How can we use large digital corpora to automatically induce dialect difference?**
- ▶ Why would “netspeak” features vary with geography?

Measuring regional specificity

Per region r ,

- ▶ **Difference** in frequencies,
 $f_{i,r} - f_i$

word	c_{SF}	c_{USA}
.	11914	907185
hella	398	3332
!	3677	276604
,	4898	382834
san	172	1654
?	2435	185482
for	2868	221142
sf	133	364
on	2523	194203
the	7901	630319

Measuring regional specificity

Per region r ,

- ▶ **Difference** in frequencies,
 $f_{i,r} - f_i$
- ▶ **Log-ratio** in frequencies,
 $\log f_{i,r} - \log f_i = \log \frac{f_{i,r}}{f_i}$

word	c_{SF}	c_{USA}
#mattomil	10	10
#bart	6	6
#davidlyons	5	5
cost=	5	5
#io14	5	5
#know14	4	4
haight	4	4
#gdc2014	4	4
#prejudices	4	4
muni	16	17

Measuring regional specificity

Per region r ,

- ▶ **Difference** in frequencies,
 $f_{i,r} - f_i$
- ▶ **Log-ratio** in frequencies,
 $\log f_{i,r} - \log f_i = \log \frac{f_{i,r}}{f_i}$
- ▶ **Regularized**
maximum-likelihood
estimate

word	c_{SF}	c_{USA}
bart	52	98
#sfgiants	56	138
francisco	91	235
sf	133	364
oakland	51	219
giants	51	334
warriors	46	368
bay	95	788
hella	398	3332
fasho	38	344

$$\hat{\eta}_r = \arg \max_{\eta} \log P(\text{counts} \mid \eta; f) - \lambda |\eta|$$

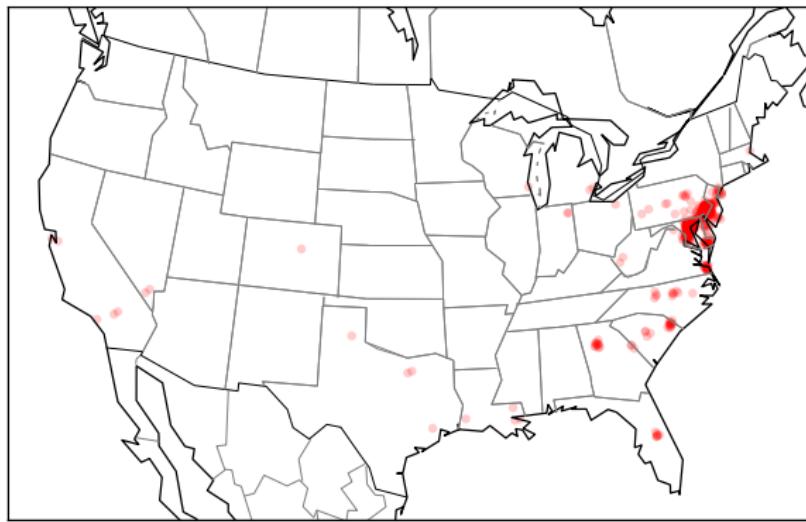
Discovered words

- ▶ **New York:** flatbush, baii, brib, bx, staten, mta, odee, soho, deadass, werd
- ▶ **Los Angeles:** pasadena, venice, anaheim, dodger, disneyland, angeles, compton, ucla, dodgers, melrose
- ▶ **Chicago:** #chicago, lbvs, chicago, blackhawks, #bears, #bulls, mfs, cubs, burbs, bogus
- ▶ **Philadelphia:** jawn, ard, #phillies, sixers, phils, wawa, philadelphia, delaware, philly, phillies

place names *entities* words

ard

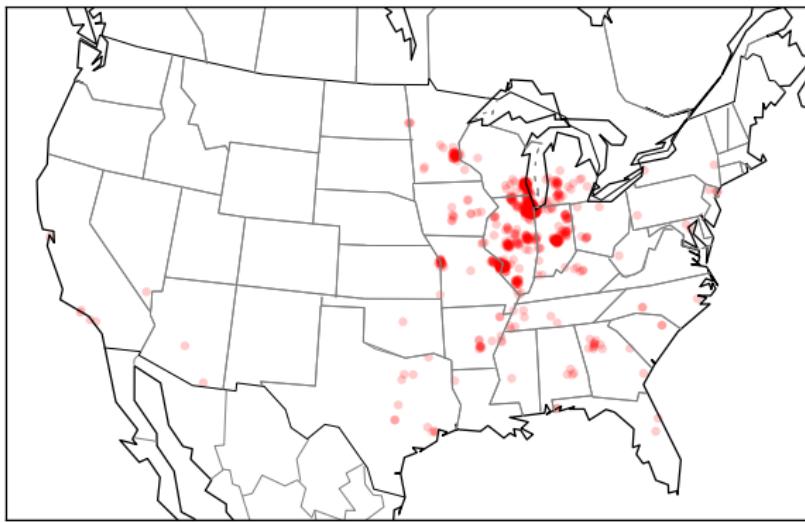
alternative spelling for alright



- ▶ @name ard let me kno
- ▶ lol u'll be ard

lbvs

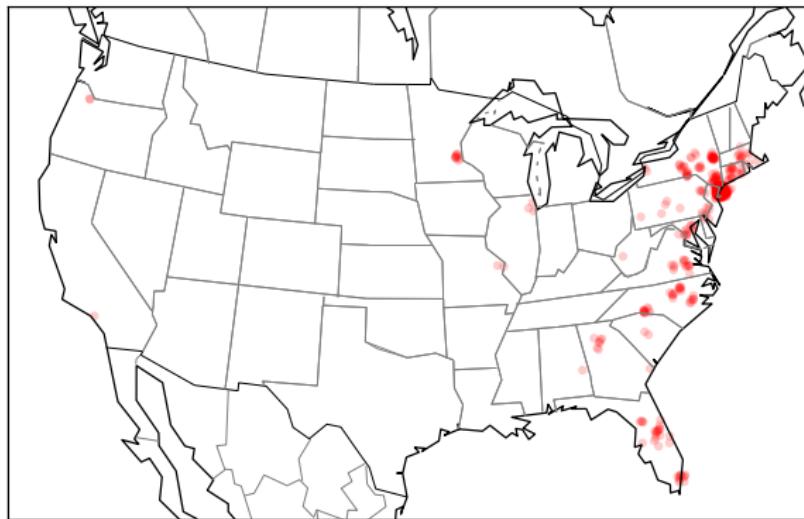
laughing but very serious



- ▶ i wanna rent a hotel room just to swim lbvs
- ▶ tell ur momma 2 buy me a car lbvs

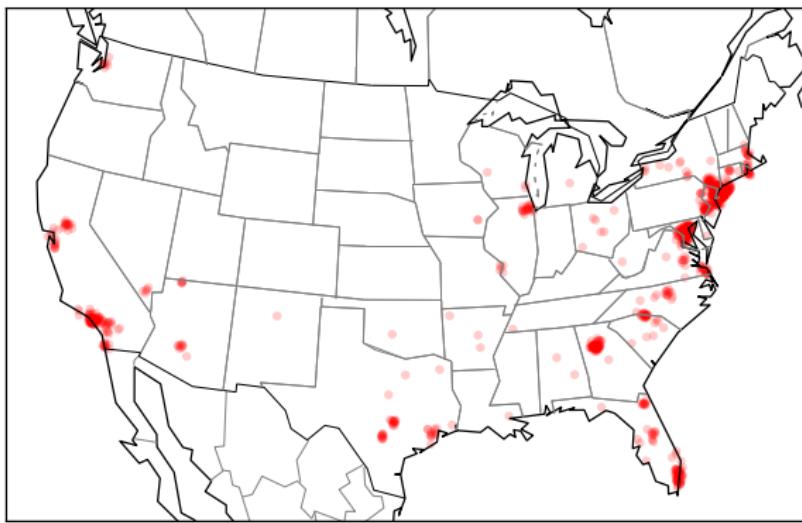
odee

intensifier, related to **overdose** or **overdone**



- ▶ i'm odee sleepy
- ▶ she said she odee miss me
- ▶ its rainin odee :(

-_- -
emoticon indicating mild annoyance



- ▶ flight delayed -_- just what i need

Variation in digital writing

Linguistic variables

- ▶ lexical items from speech

Social variables

- ▶ geography

Variation in digital writing

Linguistic variables

- ▶ lexical items from speech
- ▶ novel orthographies

Social variables

- ▶ geography

Phonologically-motivated variables

-t,-d deletion jus, ol

th-stopping dis, doe

r-lessness togetha, neva, lawd, yaself, shawty

vowels tha (the), mayne (man), bruh, brah
(bro)

relaxed pronunciations prollly, aight

“allegro spellings”⁵ gonna, finna, fitna, bouta,
tryna, iono

G-deletion



Cloyd Rivers @CloydRivers 11 Jun
Education is important, but **goin'** fishin' is importanter. Merica.
 Retweeted 2768 times
[Expand](#)  [Reply](#)  [Classic RT](#)  [Retweet](#)  [Favorite](#)  [More](#)

- ▶ In speech, “g” is deleted more often from verbs.
Does this syntactic conditioning transfer to writing?

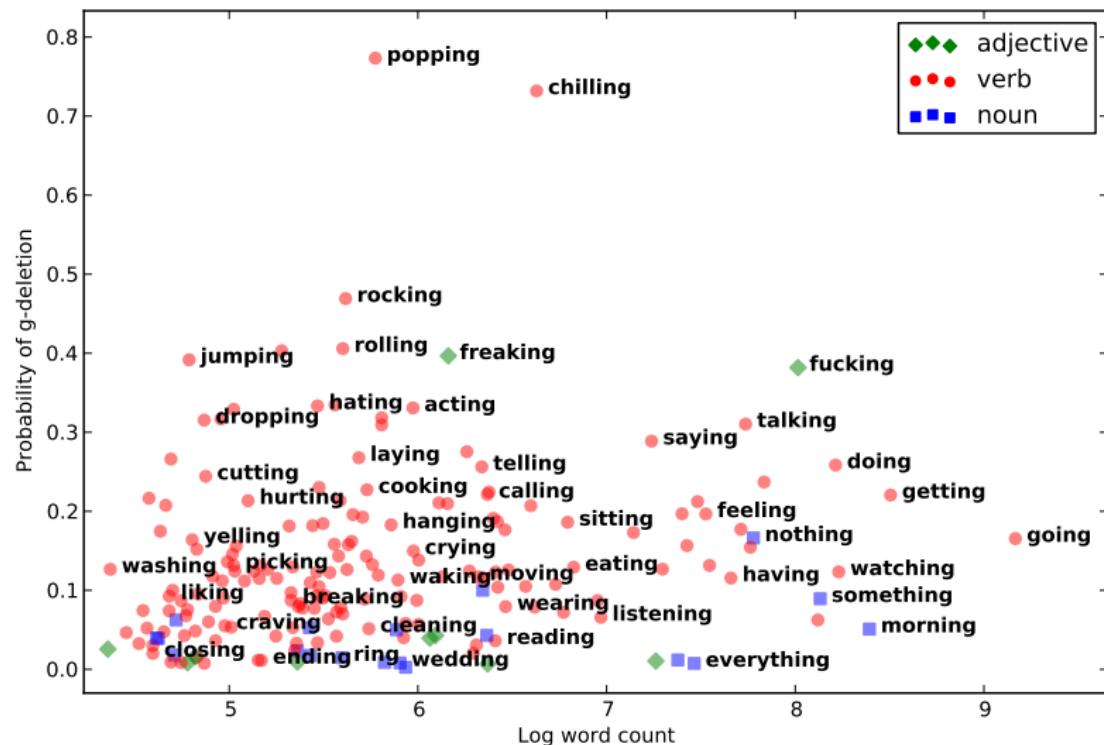
G-deletion



Cloyd Rivers @CloydRivers 11 Jun
Education is important, but goin' fishin' is importanter. Merica.
 Retweeted 2768 times
[Expand](#)     Favorite  More

- ▶ In speech, “g” is deleted more often from verbs.
- ▶ Does this syntactic conditioning transfer to writing?
- ▶ Corpus: 120K tokens of top 200 unambiguous -ing words (ex. king, thing, sing)
- ▶ Part-of-speech tags from CMU Twitter tagger.⁶

G-deletion: type-level analysis



(Colored by most common POS tag)

G-deletion: logistic regression

	Log odds	%	N
Verb	.227	.200	89,173
Noun	-.013	.083	18,756
Adjective	-.213	.149	4,964
monosyllable	-2.57	.001	108,804
Total	.178	112,893	

G-deletion: logistic regression

	Log odds	%	N
Verb	.227	.200	89,173
Noun	-.013	.083	18,756
Adjective	-.213	.149	4,964
monosyllable	-2.57	.001	108,804
High Euro-Am county	-.194	.117	28,017
High Afro-Am county	.145	.241	27,022
High pop density county	.055	.228	27,773
Low pop density county	-.017	.144	28,228
Total	.178	112,893	

Variation in digital writing

Linguistic variables

- ▶ lexical items from speech
- ▶ novel orthographies

Social variables

- ▶ geography

Variation in digital writing

Linguistic variables

- ▶ lexical items from speech
- ▶ novel orthographies
- ▶ phonetically-motivated spellings

Social variables

- ▶ geography
- ▶ demographics

Two broad categories of variables

1. Imported from speech

- ▶ Lexical variables (jawn, hella)
- ▶ Phonologically-inspired variation (-g and -t,-d deletion)
- ▶ These variables bring traces of their social and linguistic properties from speech.

2. Endogenous to digital writing

- ▶ Abbreviations (ard, lbvs, odee, ctfu, asl, ...)
- ▶ Emoticons (-__-)
- ▶ Why should these vary with geography?

Questions for this talk

- ▶ Are spoken dialect features transcribed in digitally-mediated writing?
- ▶ **How can we use large digital corpora to automatically induce dialect difference?**
- ▶ Why would “netspeak” features vary with geography?

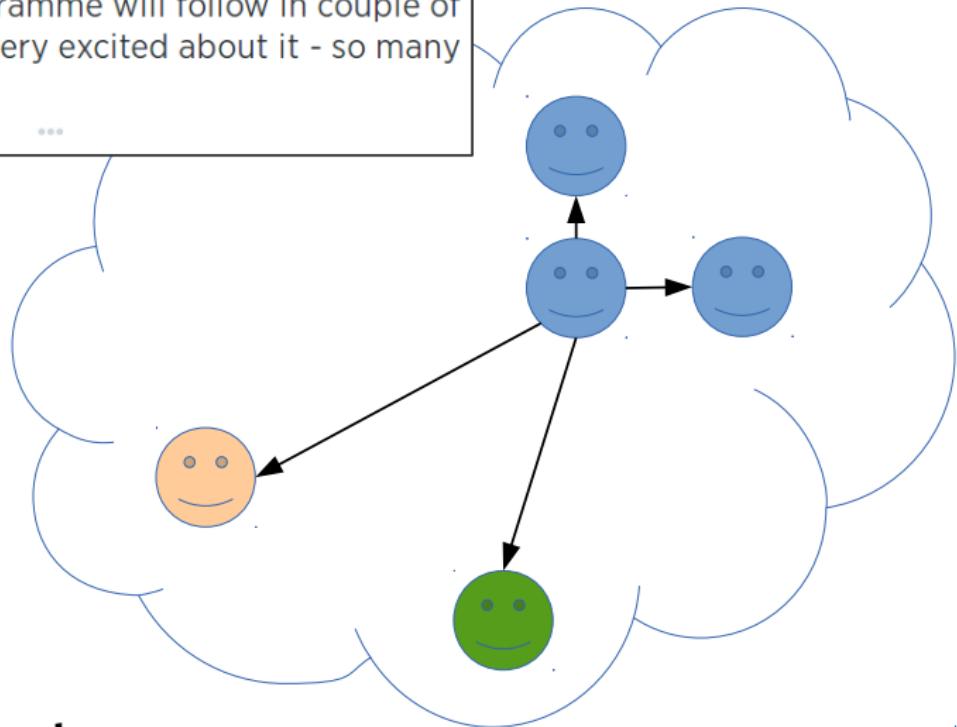
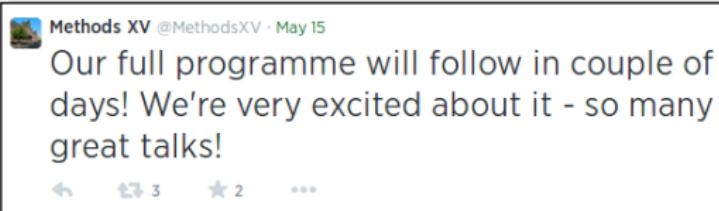
Questions for this talk

- ▶ Are spoken dialect features transcribed in digitally-mediated writing?
- ▶ How can we use large digital corpora to automatically induce dialect difference?
- ▶ **Why would “netspeak” features vary with geography?**

Language variation in a social network

- ▶ Linguistic innovations diffuse to new authors through social networks.
- ▶ In Twitter, 97% of strong ties are geographically local.
- ▶ Does this explain geographical variation in netspeak?





Broadcast



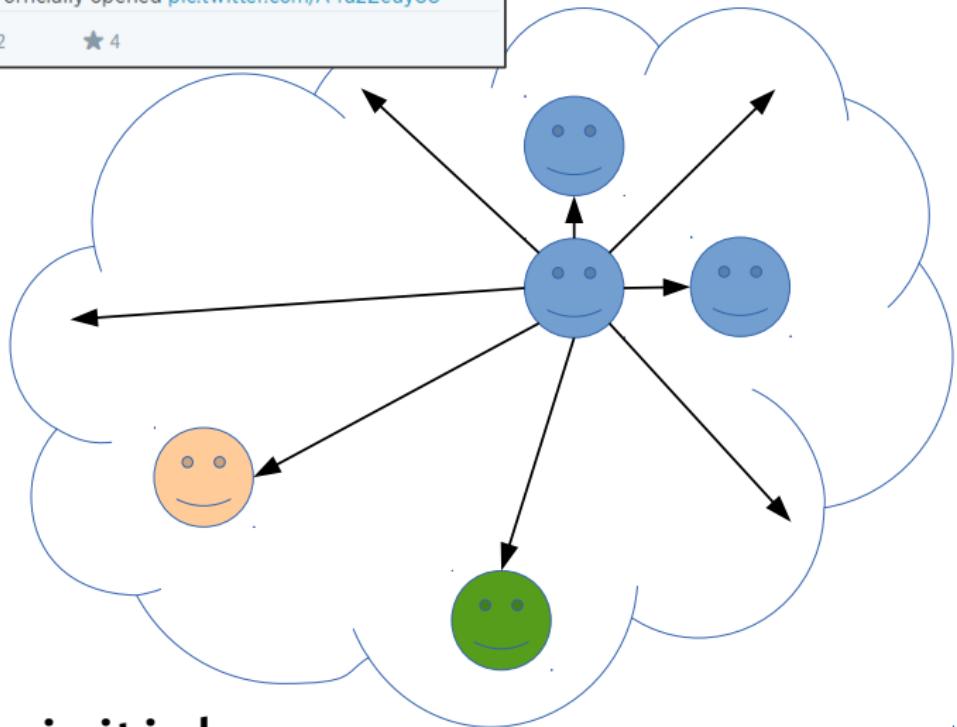
Methods XV @MethodsXV · Aug 11 · ... More

#methodsxv has officially opened pic.twitter.com/A4u2Zeuy8U



2

4



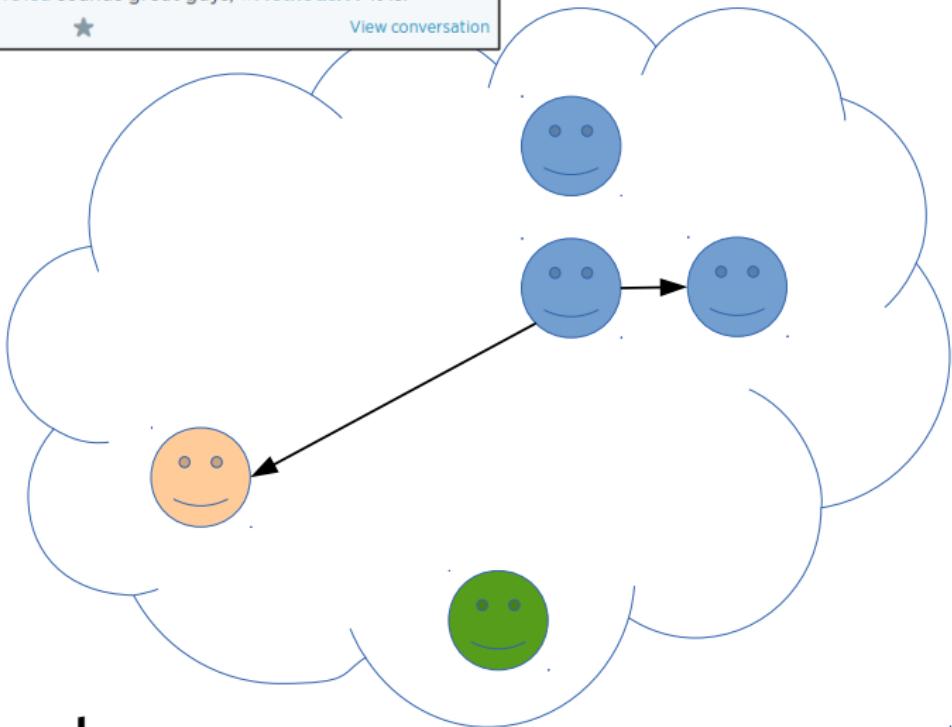
Hashtag-initial



Methods XV @MethodsXV · Aug 10 · ... More

@ajvYUL @wgi_pr3lea sounds great guys, #MethodsXV it is!

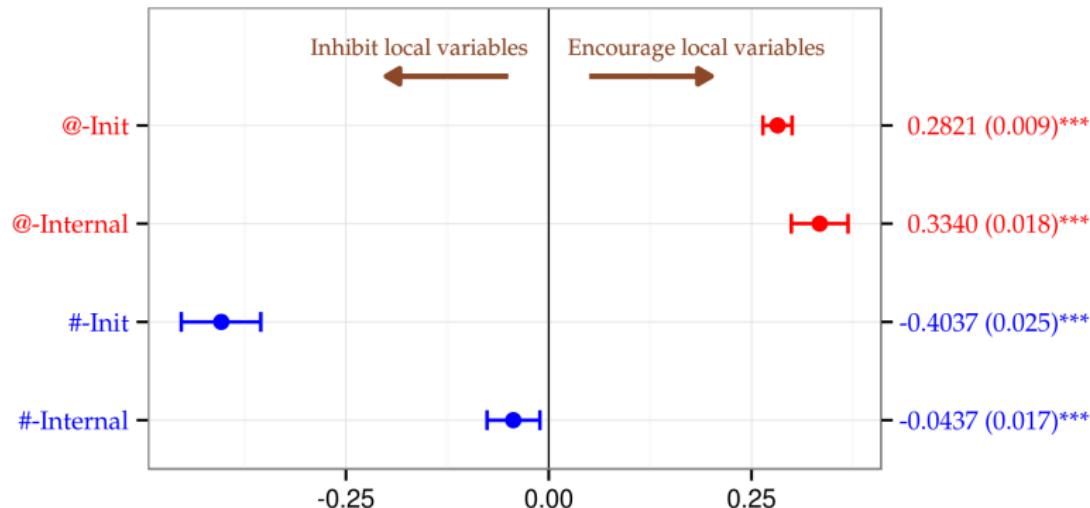
[View conversation](#)



Addressed

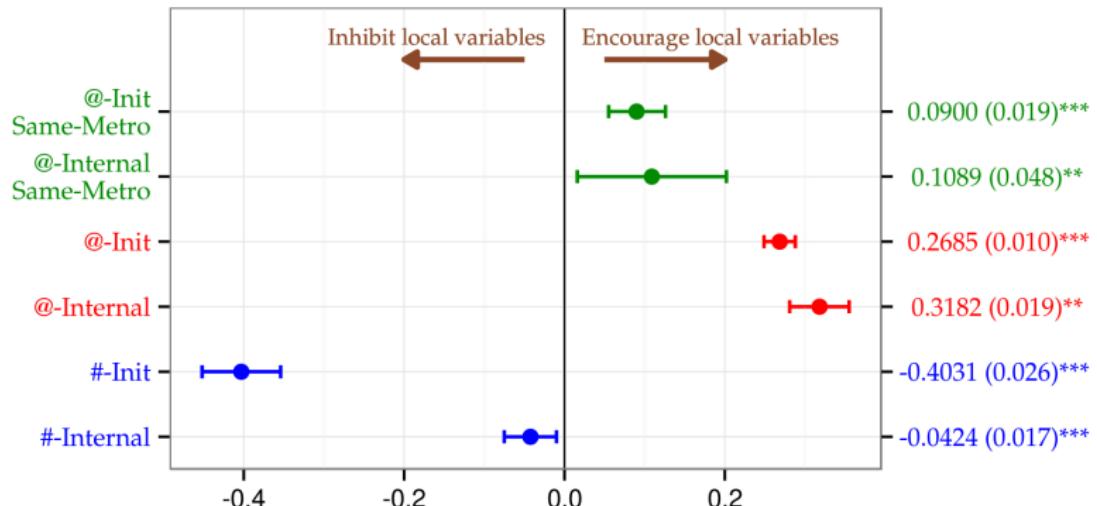
Small audience → more local language

Logistic regression from context to local netspeak:



Local audience → more local language

Logistic regression from context to local netspeak:



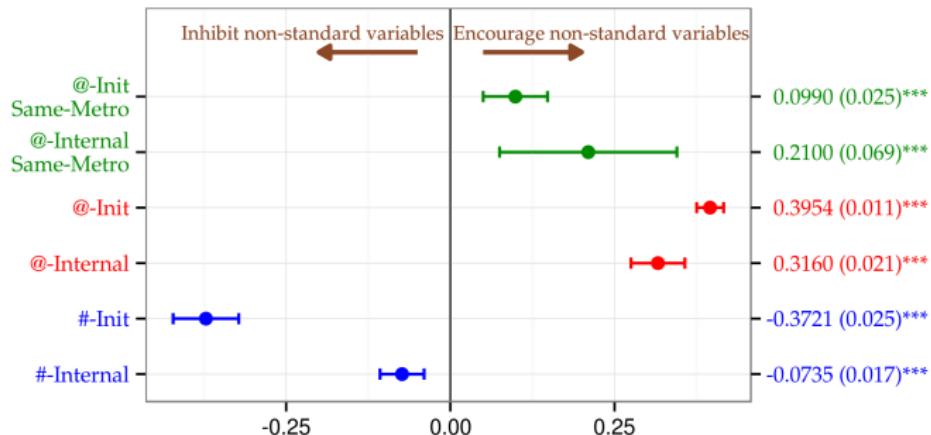
Do people know that these words are local,
and deliberately use them with local audiences?

Local audience → netspeak language?

Logistic regression from context to **non-local** netspeak:
e.g., lol, im, lmao, ya, haha ...

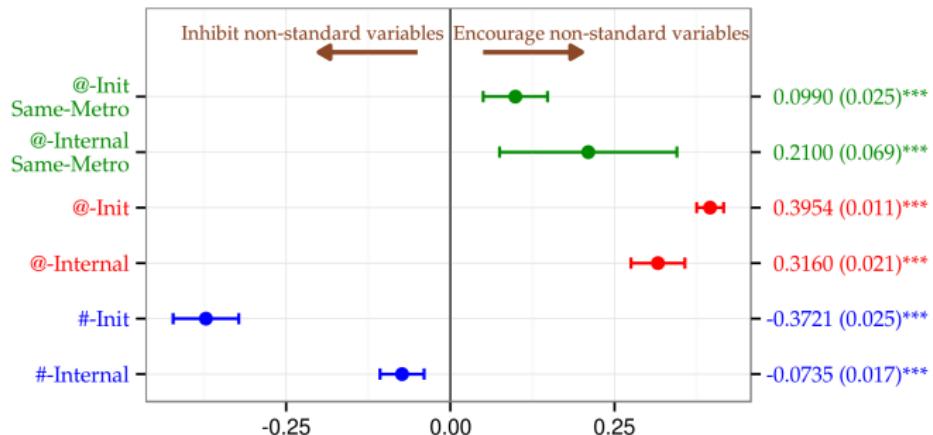
Local audience → netspeak language?

Logistic regression from context to **non-local** netspeak:
e.g., lol, im, lmao, ya, haha ...



Local audience → netspeak language?

Logistic regression from context to **non-local** netspeak:
e.g., lol, im, lmao, ya, haha ...



Non-standard words are used more often with local audiences,
even when the word itself is not local.

Questions for this talk

- ▶ Are spoken dialect features transcribed in digitally-mediated writing?
- ▶ How can we use large digital corpora to automatically induce dialect difference?
- ▶ **Why would “netspeak” features vary with geography?**

Summary

- ▶ Just as in speech, variation in digital writing is driven by a complex interplay of linguistic and social (and technological!) factors.
- ▶ The relaxed enforcement of standard language in digital media is transforming the social and communicative role of writing.



Barack Obama @BarackObama · Feb 12

Speaking of #YOLO: oaf.bo/h2sp



1.7K

2.4K

...

[View summary](#)

Summary

- ▶ Just as in speech, variation in digital writing is driven by a complex interplay of linguistic and social (and technological!) factors.
- ▶ The relaxed enforcement of standard language in digital media is transforming the social and communicative role of writing.



Barack Obama @BarackObama · Feb 12

Speaking of #YOLO: ofa.bo/h2sp



1.7K

2.4K

...

[View summary](#)

Thanks to my collaborators David Bamman, Umashanthi Pavalanathan, Tyler Schnoebelen, and to support from the National Science Foundation.

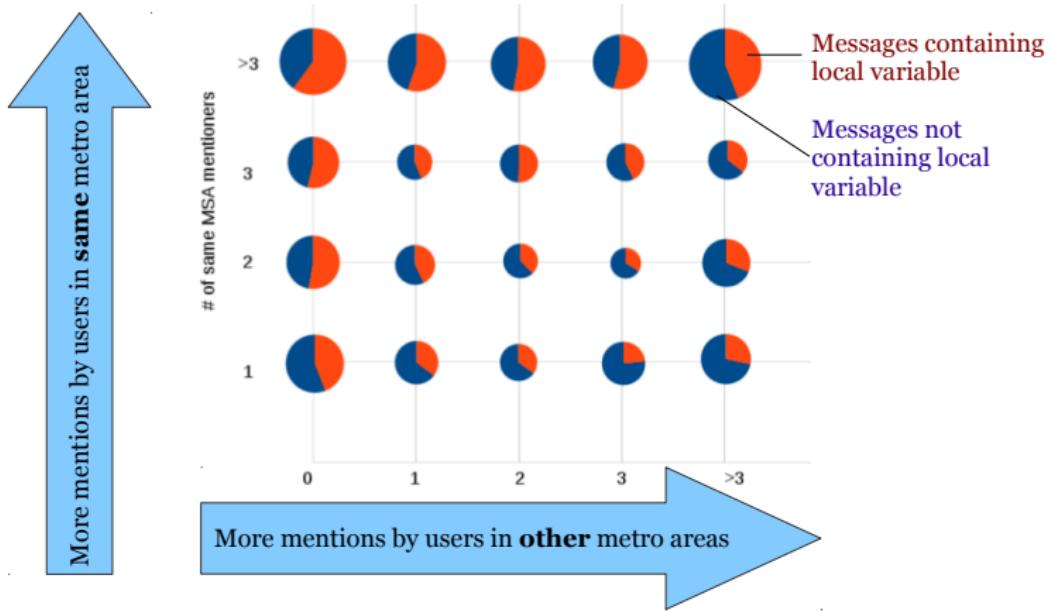
References I

- [1] Mary Bucholtz, Nancy Bermudez, Victor Fung, Lisa Edwards, and Rosalva Vargas. Hella nor cal or totally so cal? the perceptual dialectology of california. *Journal of English Linguistics*, 35(4):325–352, 2007.
- [2] H. Samy Alim. Hip hop nation language. In Alessandro Duranti, editor, *Linguistic Anthropology: A Reader*, pages 272–289. Wiley-Blackwell, Malden, MA, 2009.
- [3] Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. Sparse additive generative models of text. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1041–1048, Seattle, WA, 2011.
- [4] Jacob Eisenstein. Written dialect variation in online social media. In Charles Boberg, John Nerbonne, and Dom Watt, editors, *Handbook of Dialectology*. Wiley, 2015.
- [5] Dennis R. Preston. The Li'l Abner syndrome: Written Representations of Speech. *American Speech*, 60(4):328–336, 1985.

References II

- [6] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-of-speech tagging for Twitter: annotation, features, and experiments. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 42–47, Portland, OR, 2011.
- [7] Jacob Eisenstein. Phonological factors in social media writing. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 11–19, Atlanta, 2013.
- [8] Jacob Eisenstein. Structured variation in social media writing. *in progress*, 2015.

Local audience → more local language



Do people know that these words are local,
and deliberately use them with local audiences?

Methodological pros and cons

- ▶ Orders of magnitude more data
(enabling the study of rare linguistic phenomena
and the **induction** of unknown variables)
- ▶ Biased, non-representative population sample
- ▶ Informal communication, outside the interview
setting
- ▶ Metadata on location, time, and social
networks

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.