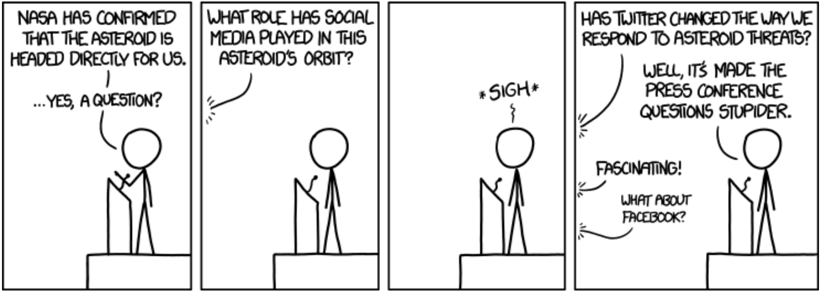
Social Media Metadata as Sociolinguistic Evidence

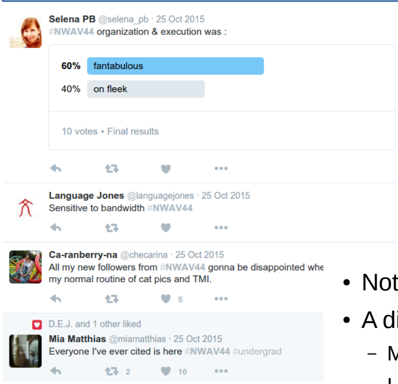
Jacob Eisenstein
@jacobeisenstein

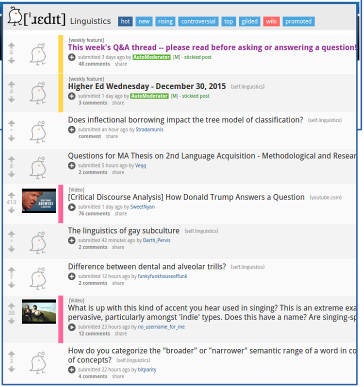
Georgia Institute of Technology

March 9, 2016

Social media





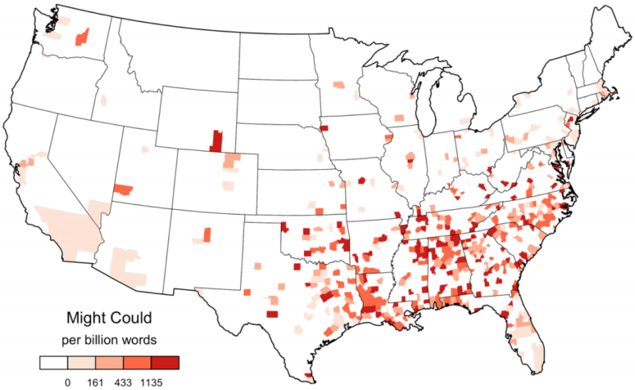


- Not a register, genre, or dialect.
- A diverse array of communicative platforms
  - Mostly text
  - Largely informal
- Several platforms support large-scale data acquisition

Why you should care about social media

- Scale
- Variation and change
- Metadata

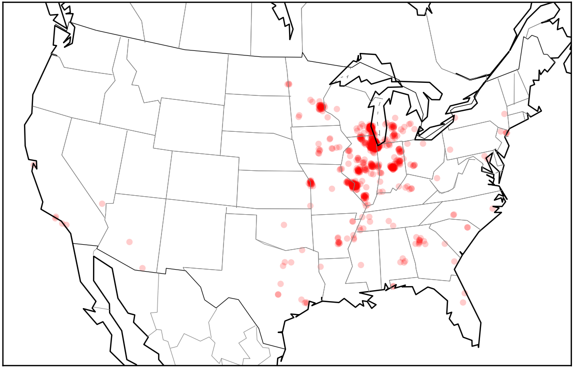
Scale: quantification of rare phenomena



(From Jack Grieve at NWAV44)

Scale: discovery of new variables

lbvs: laughing but very serious



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

(?)

Why you should care about social media

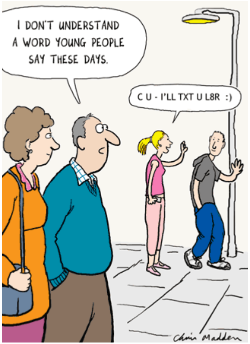
- **Scale**
- Variation and change
- Metadata

Why you should care about social media

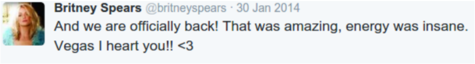
- Scale
- **Variation and change**
- Metadata

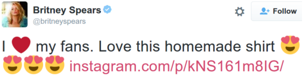
Change: Emerging norms

- In social media, writing is being used in new, speech-like contexts.
- It is constantly acquiring new affordances for paralinguistic communication:
  - from the linguistic creativity of individual users...
  - ... and from the platforms themselves.
- These affordances are highly contested! (**?**)



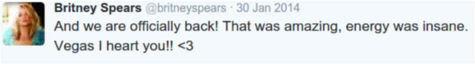
Change: Emojis versus emoticons





Does the introduction of predefined *emoji* symbols replace the functions of nonstandard orthography, such as emoticons? (?)

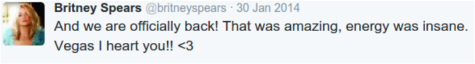
Change: Emojis versus emoticons



Does the introduction of predefined *emoji* symbols replace the functions of nonstandard orthography, such as emoticons? (**?**)

?: Emojis crowd out emoticons...

Change: Emojis versus emoticons



Does the introduction of predefined *emoji* symbols replace the functions of nonstandard orthography, such as emoticons? (**?**)

?: Emojis crowd out emoticons...

... and encourage more standard spellings!

Why you should care about social media

- Scale
- **Variation and change**
- Metadata

Why you should care about social media

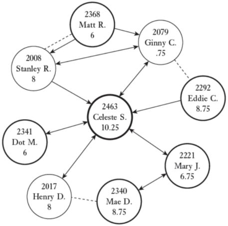
- Scale
- Variation and change
- **Metadata**
  - Social media often includes sociolinguistically relevant metadata:
    - social networks
    - demographics

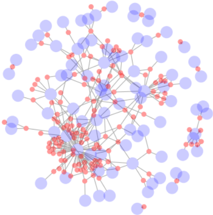
Why you should care about social media

- Scale
- Variation and change
- Metadata
  - Social media often includes sociolinguistically relevant metadata:
    - **social networks**
    - demographics

Social networks in sociolinguistics

- Reveal the microstructure of language change (**??**).
- Modulate the influence of demographic categories (**?**).
- Define local communities of linguistic practice (**?**).





Social networks in social media

Social media platforms offer a number of forms of metadata that capture social networks.

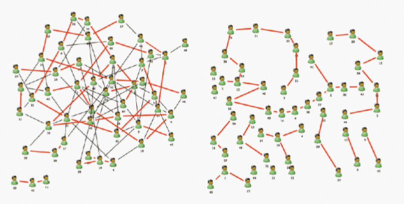
Articulated network
: Explicitly-defined connections; undirected in Facebook, directed in Twitter.

Behavioral network
: Inferred from conversational interactions, such as replies or mentions.

(?)

Social networks on Twitter

- Twitter users often follow 1000s of other users.
- Mention networks are smaller, and arguably more socially meaningful.
- Twitter query rate limiting makes mention network much easier to obtain.



(a) All links are declared followees and the red links are actual friends.

(b) After removing the black links and reorganizing the network look simpler than before. This is the hidden network that matters the most.

(?)

Case study 1: Online audience design

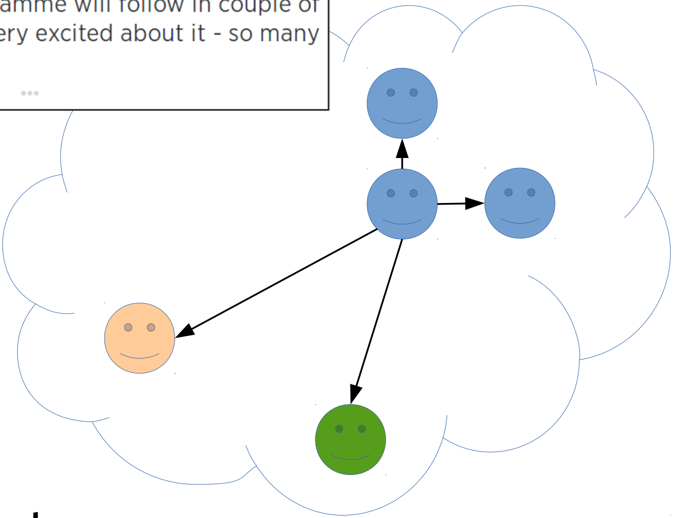
Do social media users modulate the standardness of their language according to the audience?

- Social media services offer digital affordances to control the likely audience for a message.
- The connection between language variation and these affordances can reveal the social meaning of each.
- Geographical variables (like lbvs) that are reserved for more local audiences may be more stable.

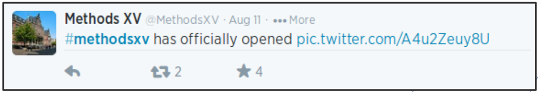
Methods XV @MethodsXV · May 15

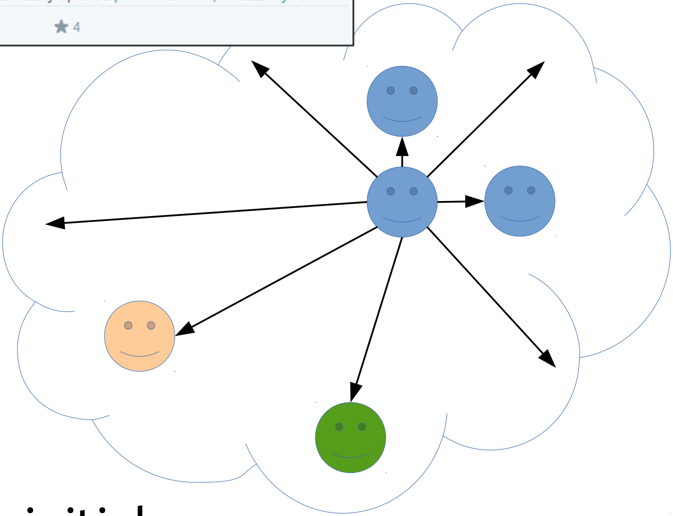
Our full programme will follow in couple of days! We're very excited about it - so many great talks!

3 2

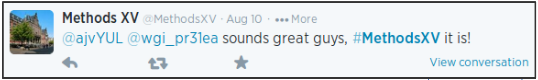


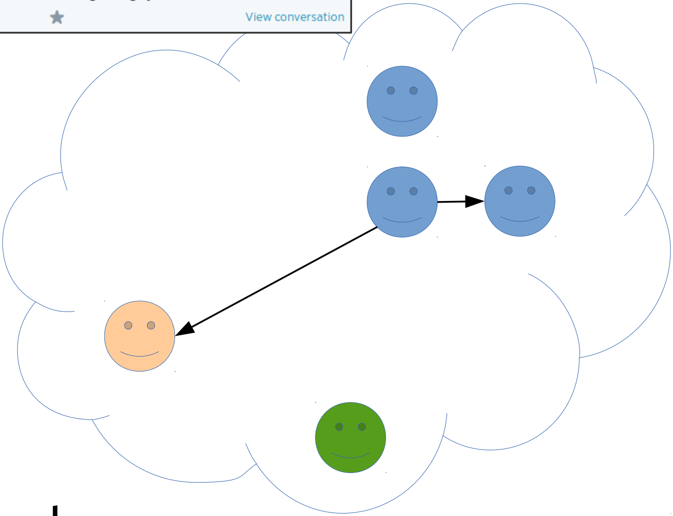
Broadcast





Hashtag-initial


Methods XV @MethodsXV · Aug 10 · ••• More
@ajvYUL @wgi_pr31ea sounds great guys, #MethodsXV it is!
View conversation

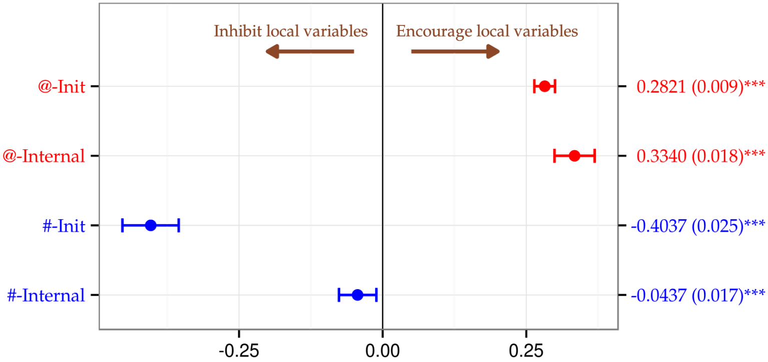


Addressed

Logistic regression

- **Dependent variable**: does the tweet contain a non-standard, geographically-specific word (e.g., lbvs, hella, jawn)
- **Predictors**
  - **Message type**: broadcast, addressed, #-initial
  - **Controls**: message length, author statistics

Small audience → less standard language



Distinguishing local ties

To distinguish **local** audiences:

- Use GPS metadata to identify author locations
- Associate metro m with user u if u is @-mentioned by:
  - at least three users within metro m;
  - nobody outside metro m.

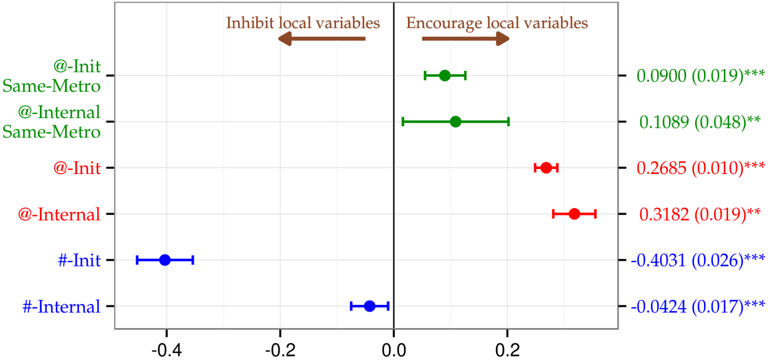
Distinguishing local ties

To distinguish **local** audiences:

- Use GPS metadata to identify author locations
- Associate metro m with user u if u is @-mentioned by:
    - at least three users within metro m;
    - nobody outside metro m.

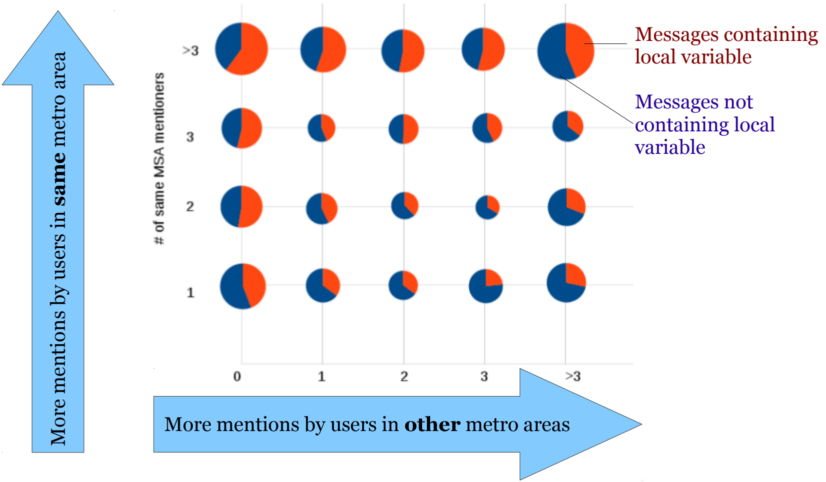
The social network lets us impute the locations of unknown users from the 1-2% of users who reveal their GPS! (?)

Local audience → less standard language



Local ties make non-standard language even more likely.

Local audience → less standard language



Why you should care about social media

- Scale
- Variation and change
- Metadata
  - Social media often includes sociolinguistically relevant metadata:
    - **social networks**
    - demographics

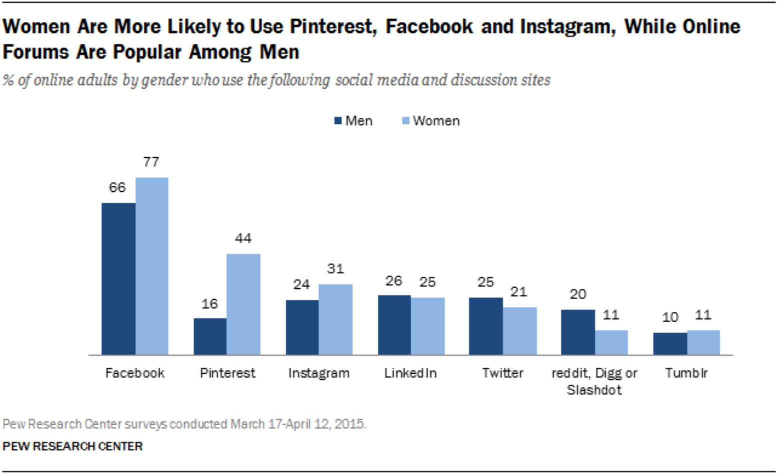
Why you should care about social media

- Scale
- Variation and change
- Metadata
  - Social media often includes sociolinguistically relevant metadata:
    - social networks
    - **demographics**

Demographics in social media

	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

Demographics in social media



Demographics from geography

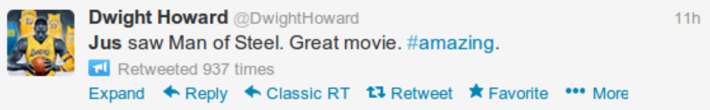
- The U.S. Census collects detailed demographics for multiple levels of geographic detail.
- For each geotagged message, treat the average census demographics as a predictor.
- Possible objections:
  - Census regions are too demographically heterogeneous.
  - People move around too much.
  - Social media users are not a representative sample of their census region.

Case study 2: Dialect in writing

Some non-standard spellings hint at dialectal pronunciations:

(ing)

(-t,-d)



How do the demographic properties of these spellings align with the demographics of the associated pronunciations? (**?**)

G-deletion

	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

(“high” / “low” = top/bottom quartile)

G-deletion

	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

("high" / "low" = top/bottom quartile)

-t,-d deletion

	Weight	Log odds	%	N
Vowel succeeding context	.483	-.066	.385	9,004
Total			.423	89,174

-t,-d deletion

	Weight	Log odds	%	N
Vowel succeeding context	.483	-.066	.385	9,004
@-message	.519	.075	.436	35,240
Total			.423	89,174

-t,-d deletion

	Weight	Log odds	%	N
Vowel succeeding context	.483	-.066	.385	9,004
@-message	.519	.075	.436	35,240
High Euro-Am county	.422	-.313	.311	19,992
High Afro-Am county	.516	.065	.508	19,854
High income county	.473	-.107	.388	20,653
Medium income county	.495	.019	.406	43,135
Low income county	.532	.127	.482	25,386
Total			.423	89,174

Dialect in writing

- Both non-standard spellings are...
  - less frequent in counties with many European Americans;
  - more frequent in counties with many African Americans.
- (ing) is more frequent in urban counties.
- (-t,-d) is more frequent in low-income counties.

Dialect in writing

- Both non-standard spellings are...
  - less frequent in counties with many European Americans;
  - more frequent in counties with many African Americans.
- (ing) is more frequent in urban counties.
- (-t,-d) is more frequent in low-income counties.

Next questions:

- Do these observations generalize to the writers themselves?
- Does linguistic systematicity vary with demographics? (**?**)

Why you should care about social media

- Scale
- Variation and change
- Metadata
  - Social media often includes sociolinguistically relevant metadata:
    - social networks
    - **demographics**

Why you should care about social media

- Scale
- Variation and change
- Metadata
  - Social media often includes sociolinguistically relevant metadata:
    - **social networks**
    - **demographics**

Case study 3: gender and social networks

We started with a painfully simple idea: train a classifier to predict author gender from language and social network features (**?**)

- Prior work shows that text predicts author gender (**?**)
- Social networks are often assortative with respect to gender (**?**).
- Can we build a better classifier by putting these two features together?

Data

14,464 Twitter users from 2011 (56% male)

- geolocation in USA
- must use 50 of 1000 most frequent words
- no more than 1000 follow connections

In total: 9.2M tweets, from January to June 2011

Demographics from names

The U.S. Social Security Administration collects yearly statistics on given names and gender.

$$P(\text{age}, \text{gender} \mid \text{name}) = \frac{\text{count}(\text{age}, \text{gender}, \text{name})}{\text{count}(\text{name})}$$

Demographics from names

- Limited to names that occur at least 1000 times in the Census data.
  - $\sim 9,000$ names in total
  - Most infrequent: Cherylann, Kailin, Zeno
- The median author's name is 99.6% homogeneous by gender.
- 95% of all authors have a name that is at least 85% associated with one gender.

Social network

Behavioral network induced from **mutual** @-mentions

- Mentions must occur over a period of at least two weeks.
- Moderate gender assortativity:
  - Women have 58% female friends.
  - Men have 67% male friends.

Automatic classification

Logistic regression from bag-of-words features gives 88% accuracy.

- This is similar to prior work (**??**).
- Will social network homophily help fix the remaining errors?

Adding social network features

Logistic regression, 10-fold cross-validation:

- Text alone: 88% accuracy

Adding social network features

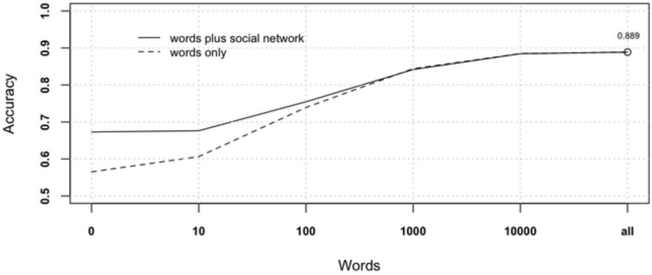
Logistic regression, 10-fold cross-validation:

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

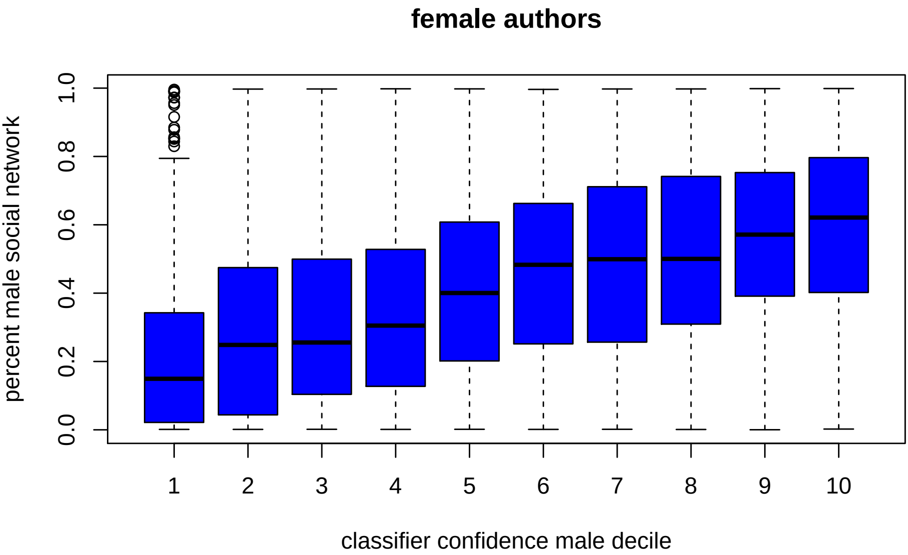
Adding social network features

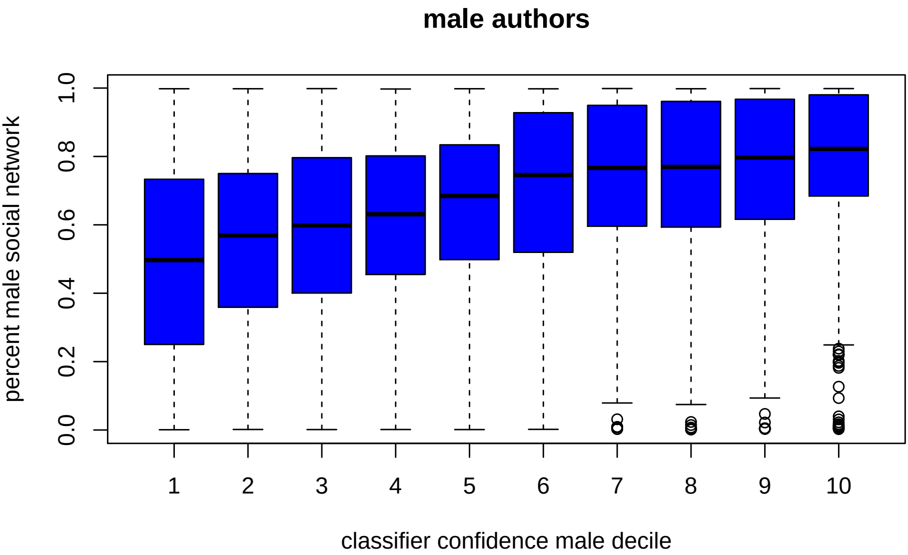
Logistic regression, 10-fold cross-validation:

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



With $\geq$ 1000 words per author, adding network info does not improve accuracy. **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

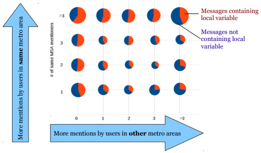
	text vs network correlation
female authors	0.38 ($.35 \leq r \leq .40$)
male authors	0.33 ($.30 \leq r \leq .36$)

Why social network features don't help

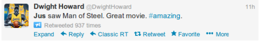
	text vs network correlation
female authors	0.38 ($.35 \leq r \leq .40$)
male authors	0.33 ($.30 \leq r \leq .36$)

- Language and social network are correlated even after controlling for author gender.
- Rather than seeing linguistic features as revealing the author's gender, they reveal an attitude towards gender.

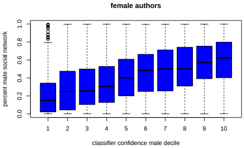
Summary of case studies



1. Audience design on the Twitter social network



2. Demographic profiles of spelling variables



3. Linking language, gender, and social networks