Computational Models of Language Change in Online Social Media

Jacob Eisenstein
@jacobeisenstein

Georgia Institute of Technology

May 11, 2016

sensei
@fauntleroy

Follow

@jtimberlake this #cantstopthefeeling is fuckin #jammin | #basslinecrazytoo

RETWEETS	LIKES
19	50

⛩ **sensei** @fauntleroy

Follow

@jtimberlake this #cantstopthefeeling is fuckin #jammin | #basslinecrazytoo

RETWEETS	LIKES
19	50

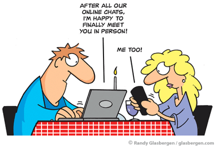
Justin Timberlake @jtimberlake · May 9

Thank you, broseph!!! 🙏‼️

> ⛩ **sensei** @fauntleroy
>
> @jtimberlake this #cantstopthefeeling is fuckin #jammin | #basslinecrazytoo

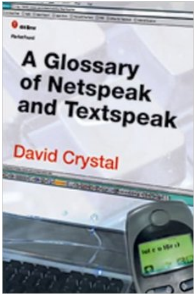
303 1.6K

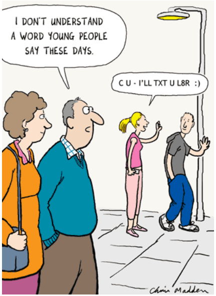
A new role for writing

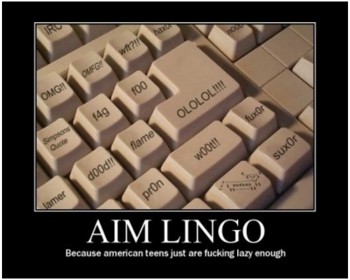


Digitally-mediated writing plays an increasingly ubiquitous role in informal communication.

"Netspeak": a new language variety?





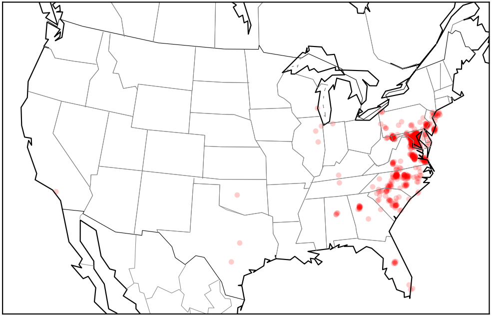


- These phenomena lead some to herald the birth of a new "netspeak" variety (Crystal, 2006).
- But is netspeak a coherent genre? A register? A dialect?

Intersections of genre and dialect

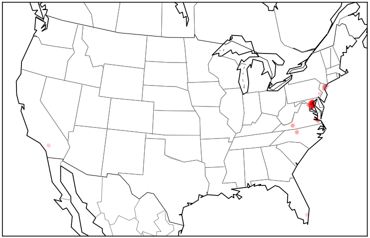
lls (laughing like shit)

- `@user lls, we need to skate homie`
- `@user lls chill`

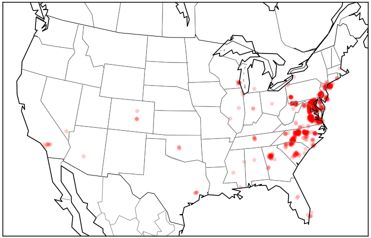


Change over time

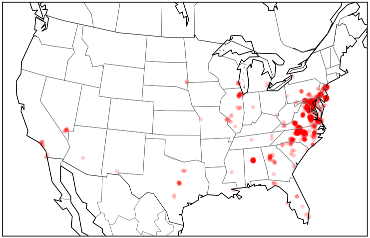
2009



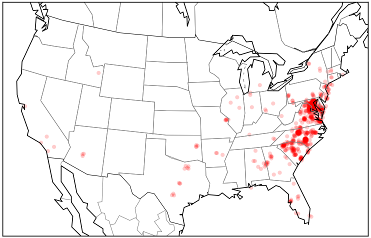
2010



2011



2012



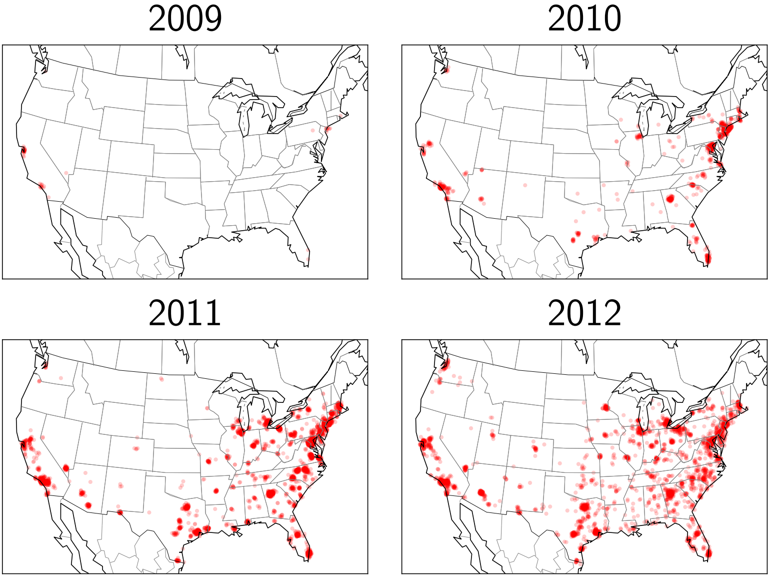
Language change as social influence

For a word like lls to spread, two things must happen:

- **Exposure**: individuals must see or hear the word in use.
- **Influence**: they must decide to use it themselves.

Language change can tell us a lot about social structures: who talks to whom, which groups tend to differentiate their linguistic features from each other, and who leads and who follows.

Possible endpoints: total victory



Possible endpoints: failure

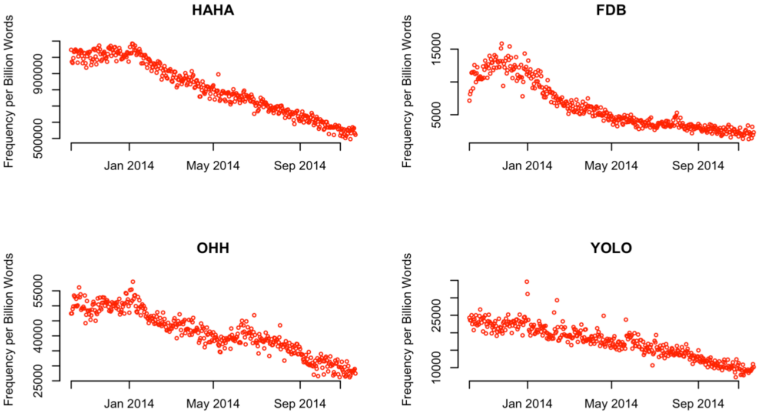
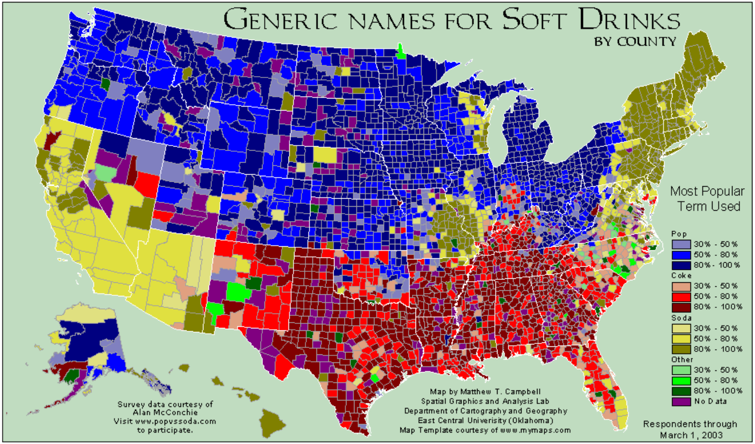
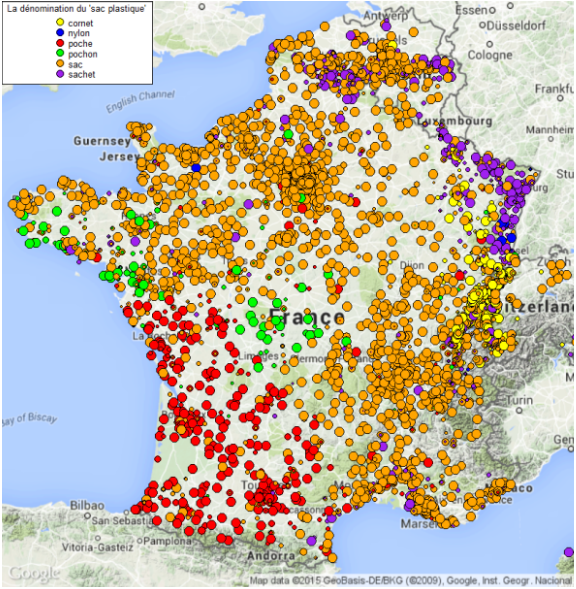


Image credit: Jack Grieve

Possible endpoints: stable variation



Possible endpoints: stable variation

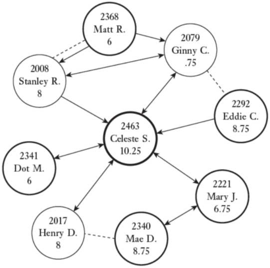


Social theories of language change

- **Cultural capital**: dialects are a form of social differentiation, which language change helps to maintain (Bourdieu, 1984).
- **Covert prestige**: stigmatized linguistic forms can convey "covert" social advantages, leading to resistance to change (Trudgill, 1972).
- **Indexicality**: non-standard forms "index" various social attributes, and speakers creatively draw from these social associations to craft distinct personal styles (Eckert, 2008).

Sociolinguistic approaches

- Sociolinguistics relies heavily on the method of **apparent time** to understand change.
- Social networks are recovered by snowball sampling and interviews.
- These methods have yielded many insights, but generalization is limited by the high cost of data acquisition.



Large-scale study of language change

Social media analysis offers several advantages:

- **Scale**: by studying millions of speakers, it is possible to make more confident generalizations and to investigate more rare phenomena.
- **Speed**: language change in online social media is currently so rapid that real time investigation is practical.
- **Metadata**: explicit records of social interactions make it far more feasible to link language with social structures.

Hypotheses

This work uses large-scale Twitter data to test main hypotheses about language change.

- **H1**: language change is transmitted across social networks that are visible from metadata in online social media platforms.
- **H2**: geographically local social network ties are better conduits of language change.
- **H3**: strong ties (densely embedded) are better conduits of language change.

Dataset

- Twitter analysis is usually conducted on a **sample** from the streaming API (e.g., Eisenstein et al., 2010, 2014).
- But modeling the fine structure of language change requires **complete data**, because random samples will miss most of the co-occurrences that reveal sociolinguistic influence.
- This work: a dataset of all public tweets between 2011 and 2014, with 4.35 million unique user accounts.

Geography

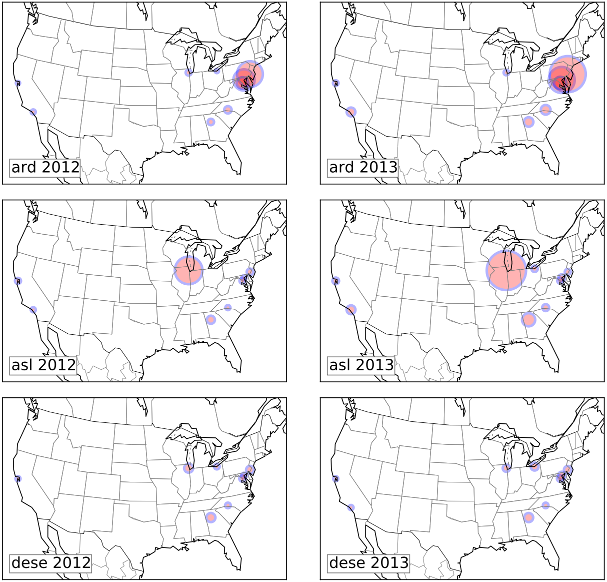
We focus on nine metropolitan areas in the USA:

- Atlanta
- Baltimore
- Charlotte
- Chicago
- Cleveland
- Los Angeles
- Philadelphia
- San Francisco
- Washington, DC

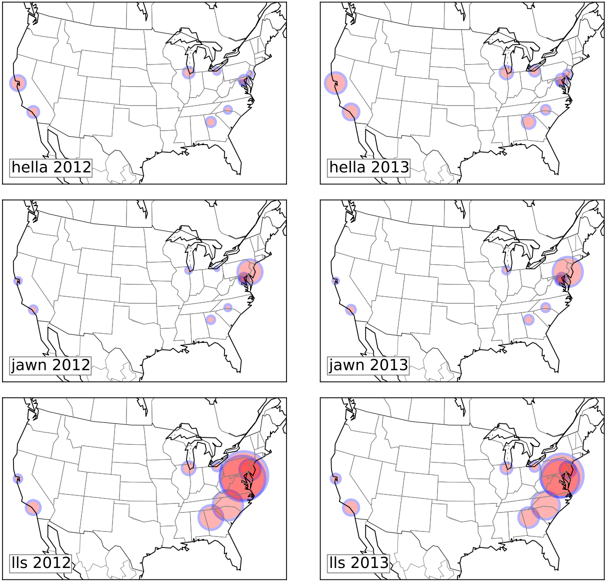
Linguistic variables

	count	type	definition
ard	28,882	phonetic	alternative spelling for alright, e.g., lol ard ill text u
asl	36,159	abbreviation	intensifier, e.g., I'm hungry asl
dese	1,664	phonetic	alternative spelling for these, e.g., I ain't like sum of dese frauds
hella	20,470	lexical	intensifier, e.g., I'm hella hungry
jawn	14,416	lexical	generic noun, e.g., that's my jawn
lls	317,403	abbreviation	laughing like shit, e.g., lls i wish; stay mad lls

Linguistic variables



Linguistic variables



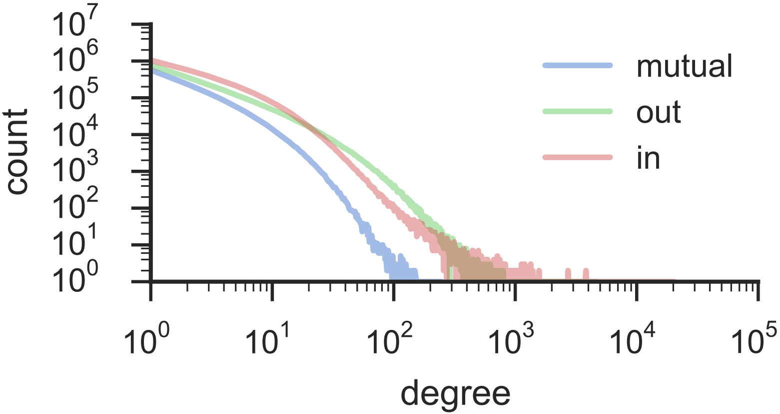
Social network

Two users are considered to have a social network tie if they have each mentioned each other in a message, e.g.

- User1: @user2 salut
- User2: @user1 what's up?

This "mention network" has been argued to be more descriptive of meaningful social ties than the "articulated network" of follower-followee links (Huberman et al., 2008; Puniyani et al., 2010).

Social network



The symmetrized ("mutual") mention network yields a more credible degree distribution.

Summary of data

Social network

Bart	Lisa
Bart	Milhouse
Lisa	Homer
Homer	Barney
...	...

Locations

Bart	Los Angeles
Milhouse	Los Angeles
Lisa	Atlanta
Homer	Chicago
...	...

Language

Bart	jawn	Feb 1, 2013, 13:45
Milhouse	jawn	Feb 1, 2013, 13:50
Homer	hella	Feb 1, 2013, 18:15
Bart	lls	Feb 2, 2013, 07:30
Milhouse	lls	Feb 2, 2013, 07:40
...	...	...

Point processes

Hypotheses

- **H1**: language change is transmitted across social networks that are visible from metadata in online social media platforms.
- **H2**: geographically local social network ties are better conduits of language change.
- **H3**: strong ties (densely embedded) are better conduits of language change.

We test hypotheses using **point process models**, which are generative models over **event cascades**.

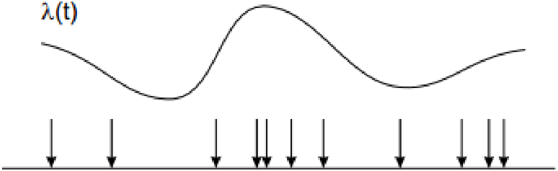
The Poisson process

- Suppose we have a cascade of event times, $\{t_n\}_{n \in 1 \ldots N}$.
- Let $y(t_1, t_2)$ be the count of events between times t_1 and t_2. Then,

$$y(t_1, t_2) \sim \text{Poisson}(\Lambda(t_1, t_2)) \tag{1}$$

$$\Lambda(t_1, t_2) = \int_{t_1}^{t_2} \lambda(t) dt. \tag{2}$$

The Poisson process



For example:

- $y(t_1, t_2)$ is the count of the word lls between 2013 and 2014
- $\lambda(t)$ is the (continuously varying) intensity function.

Hawkes process

A Poisson process in which the intensity function depends on the history (Hawkes, 1971)

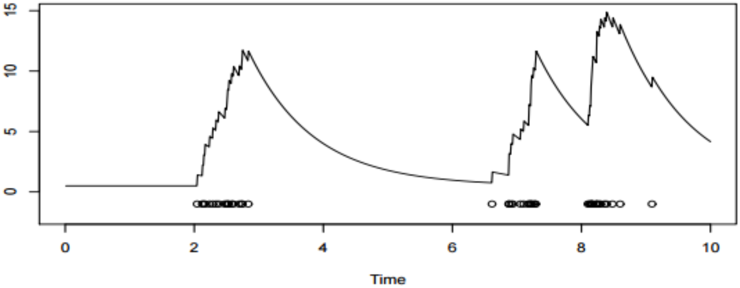
$$\lambda(t) = \mu + \alpha \sum_{t_n < t} \kappa(t - t_n), \tag{3}$$

where the time kernel κ is typically defined as,

$$\kappa(\Delta t) = e^{-\gamma \Delta t}. \tag{4}$$

- μ is the **base rate**;
- α captures the degree of self-excitation;
- γ is the time scale.

Hawkes process



For example:

- $y(t_1, y_2)$ is the count of the word lls
- α captures the tendency of usages of lls to "excite" other usages.

Multivariate Hawkes process

Now suppose each event has some *source* m.

- The cascade is $\{(t_n, m_n)\}_{n \in 1 \dots N}$.
- The intensity for source m is,

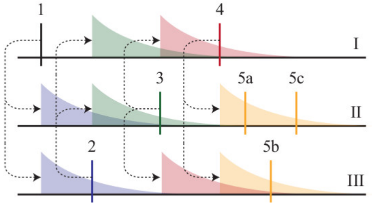
$$\lambda_m(t) = \mu_m + \sum_{t_n < t} \alpha_{m_n \to m} \kappa(t - t_n), \qquad (5)$$

where $\alpha_{m_n \to m}$ is the excitation exerted by events with source m_n on source m.

Multivariate Hawkes process

For example:

- Each source m corresponds to an individual social media user.
- $y_m(t_1, t_2)$ is the count of usages of lls by user m between t_1 and t_2.
- $\alpha_{m_1 \to m_2}$ captures the influence of m_1 on m_2.



See Blundell et al. (2012) for another application to language data.

Maximum likelihood estimation

$$\mathcal{L}(\{(t_n, m_n)\}_{n\in 1\ldots N}) = \sum_{n=1}^{N} \log \lambda_m(t_n) - \sum_{m=1}^{M} \Lambda_m(0, T) \tag{6}$$

$$= \sum_{n=1}^{N} \log \lambda_m(t_n) - \sum_{m=1}^{M} \int_0^T \lambda_m(t) dt \tag{7}$$

Estimation: maximum likelihood s.t. $\alpha > 0, \mu > 0$.

- Convex in the parameters α and μ
- Linear complexity in the number of parameters and the number events.

Maximum likelihood estimation

$$\mathcal{L}(\{(t_n, m_n)\}_{n\in 1\dots N}) = \sum_{n=1}^{N} \log \lambda_m(t_n) - \sum_{m=1}^{M} \Lambda_m(0, T) \qquad (6)$$

$$= \sum_{n=1}^{N} \log \lambda_m(t_n) - \sum_{m=1}^{M} \int_0^T \lambda_m(t)dt \qquad (7)$$

Estimation: maximum likelihood s.t. $\alpha > 0, \mu > 0$.

- Convex in the parameters α and μ
- Linear complexity in the number of parameters and the number events.

But! The number of parameters is **quadratic** in the number of sources.

Parametric Hawkes process

Let's make the infection parameters a function of shared features of each pair of individuals,

$$\alpha_{m_1 \to m_2} = \boldsymbol{\theta}^\top \boldsymbol{f}(m_1 \to m_2). \tag{8}$$

- We now need estimate only $\#|\theta|$ parameters, rather than M^2.
- Because α is an affine function of θ, convexity is preserved.
- Given binary features, non-negativity constraints on the weights $\theta_i \geq 0$ ensure that $\alpha_{m_1,m_2} \geq 0$.

Features

Self-excitation $f_1(m_1 \to m_2) = 1$ if $m_1 = m_2$, zero otherwise

Social network $f_2(m_1 \to m_2) = 1$ if there is an edge between m_1 and m_2 in the articulated social network, $(m_1, m_2) \in E$.

Locality $f_3(m_1 \to m_2) = 1$ if $(m_1, m_2) \in E$ **and** m_1 and m_2 are geolocated to the same metropolitan area.

Tie strength $f_4(m_1 \to m_2) = 1$ if $(m_1, m_2) \in E$ **and** m_1 and m_2 is a densely embedded tie.

Features

Self-excitation $f_1(m_1 \to m_2) = 1$ if $m_1 = m_2$, zero otherwise

Social network $f_2(m_1 \to m_2) = 1$ if there is an edge between m_1 and m_2 in the articulated social network, $(m_1, m_2) \in E$.

Locality $f_3(m_1 \to m_2) = 1$ if $(m_1, m_2) \in E$ **and** m_1 and m_2 are geolocated to the same metropolitan area.

Tie strength $f_4(m_1 \to m_2) = 1$ if $(m_1, m_2) \in E$ **and** m_1 and m_2 is a densely embedded tie.

Features

Self-excitation $f_1(m_1 \to m_2) = 1$ if $m_1 = m_2$, zero otherwise

Social network $f_2(m_1 \to m_2) = 1$ if there is an edge between m_1 and m_2 in the articulated social network, $(m_1, m_2) \in E$.

Locality $f_3(m_1 \to m_2) = 1$ if $(m_1, m_2) \in E$ **and** m_1 and m_2 are geolocated to the same metropolitan area.

Tie strength $f_4(m_1 \to m_2) = 1$ if $(m_1, m_2) \in E$ **and** m_1 and m_2 is a densely embedded tie.

Features

Self-excitation $f_1(m_1 \to m_2) = 1$ if $m_1 = m_2$, zero otherwise

Social network $f_2(m_1 \to m_2) = 1$ if there is an edge between m_1 and m_2 in the articulated social network, $(m_1, m_2) \in E$.

Locality $f_3(m_1 \to m_2) = 1$ if $(m_1, m_2) \in E$ **and** m_1 and m_2 are geolocated to the same metropolitan area.

Tie strength $f_4(m_1 \to m_2) = 1$ if $(m_1, m_2) \in E$ **and** m_1 and m_2 is a densely embedded tie.

Measuring tie strength

Mutual friends

$$mf(i,j) = \#|\{k : k \in \Gamma(i) \cap \Gamma(j)\}| \tag{9}$$

Measuring tie strength

Mutual friends

$$mf(i,j) = \#|\{k : k \in \Gamma(i) \cap \Gamma(j)\}| \tag{9}$$

Adamic & Adar (2003): reweight each mutual friend by its degree:

$$aa(i,j) = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log \#|\Gamma(k)|} \tag{10}$$

We set $f_4(m_1, m_2) = 1$ if $aa(m_1, m_2)$ is in the 90th percentile.

Hypothesis testing

We compare a series of **nested models**.

- $F2 + F1$ **vs** $F1$: is language change transmitted across the social network?
- **All features vs** $F1 + F2 + F4$: are local ties better conduits of language change?
- **All features vs** $F1 + F2 + F3$: are densely embedded ties better conduits of language chnage?

Each comparison is performed using a likelihood ratio test, with correction for multiple comparisons (Benjamini & Hochberg, 1995).

Results

	H1: social network	H2: local ties	H3: tie strength
ard	✓		✓
asl	✓	✓	✓
dese	✓		
hella	✓		✓
jawn	✓	✓	✓
lls	✓		✓
total	6/6	2/6	5/6

A checkmark ✓ indicates that the more complex model was significantly better at $p < .05$.

Discussion

The social network matters. Users of Twitter tend to use the same neologisms as their friends.

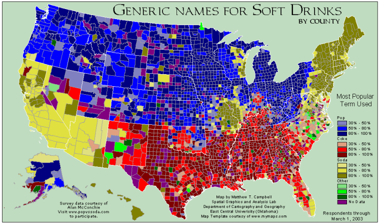
Social evaluation affects language change. Strong ties are better conduits of language change than weak ties.

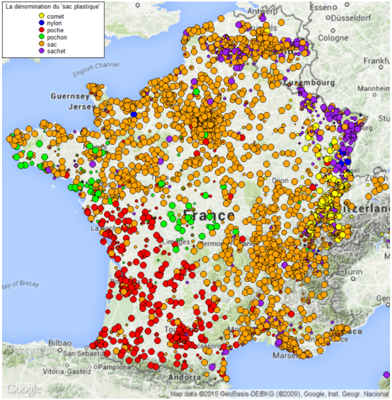
Geography's role is less clear. The evidence that Twitter users pay special attention to geographically-local ties is weak.

These findings are closely matched by comparisons of predictive likelihood on out-of-sample data.

Any hope for stable variation?

From these results, it may seem unlikely that social media writing will yield new stable geographical variables like coke/pop or sac/poche.





Exposure's geographical bias

However, *mere exposure* has a strong bias towards geographical differentiation.

- Most social network ties are local (Backstrom et al., 2010, inter alia).
- Nonstandard forms are used more frequently in messages to local friends (Pavalanathan & Eisenstein, 2015).

Even if readers do not distinguish local exposures, most exposures to neologisms are local, and this works to preserve geographical distinctions.

Next steps for this model

- More words, better aggregation across words, comparisons between groups of words.
- Use parametric Hawkes Process to compare various procedures for social network construction.
- Check whether simulations from these models qualitatively match observations.
- Control for exogenous factors, using search query records.

Bigger questions for the future

- Is the diffusion of language change different from other forms of social influence (hashtags, URLs)? Are these differences measurable in the parameters of the Hawkes process?
- Is linguistic influence correlated with other forms of influence (social, cultural, political)?

Summary

- Written language is undergoing a period of unprecedented change in online social media.
- Social media data makes it possible to study the micro-structure of language change.
- Language change can be modeled as a point process, with the use of new variable forms as a cascade of timestamped events.
- The Parametric Hawkes process enables the comparison of hypothesized social factors in the generation of these event cascades.

Thanks

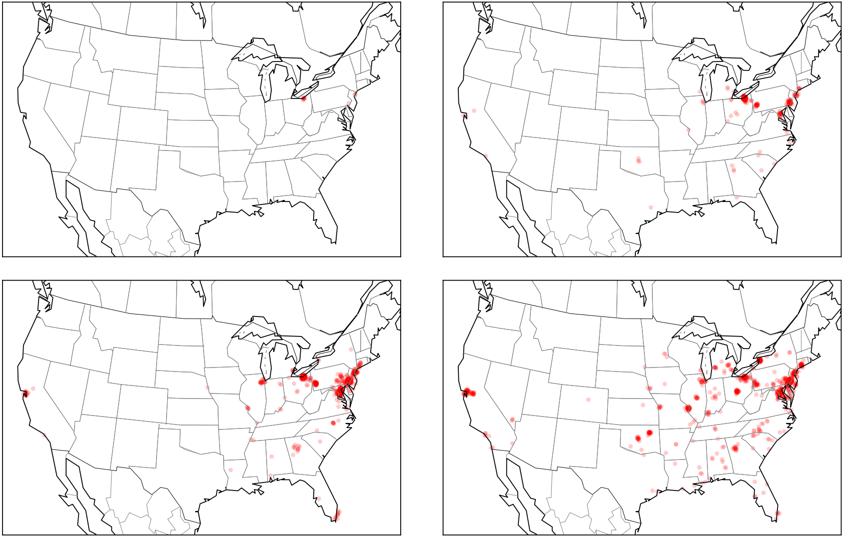
Co-authors and collaborators

- **Georgia Tech**: Rahul Goel, Sandeep Soni, Naman Goyal, Le Song
- **Columbia**: John Paparrizos
- **Microsoft Research**: Hanna Wallach, Fernando Diaz

Sponsors: National Science Foundation, National Institutes for Health, and the Air Force of Scientific Research.

Change from 2009-2012: ctfu

@name lmao! haahhaa ctfu!



The voyage of ctfu

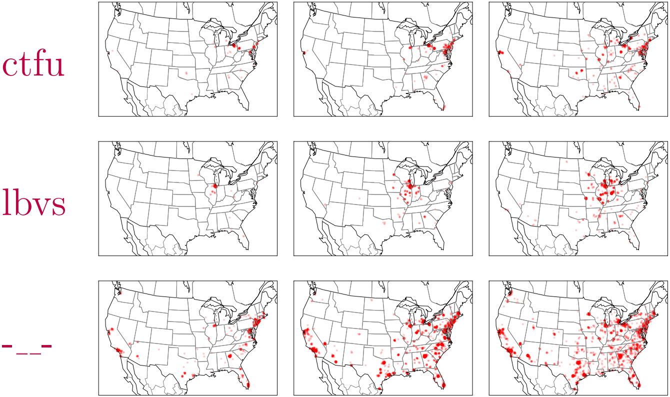
2009	Cleveland
2010	Pittsburgh, Philadelphia
2011	Washington DC, Chicago, NY
2012	San Francisco, Columbus

The voyage of ctfu

2009	Cleveland
2010	Pittsburgh, Philadelphia
2011	Washington DC, Chicago, NY
2012	San Francisco, Columbus

This trajectory is hard to explain with models based only on geography or population.

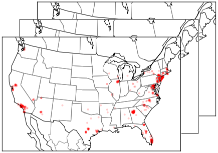
An aggregate model of lexical diffusion



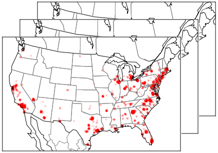
- Thousands of words have changing frequencies.
- Each spatiotemporal trajectory is idiosyncratic.
- What's the aggregate picture?

Language change as an autoregressive process

Word counts are binned into 200 metro areas and 165 weeks.



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



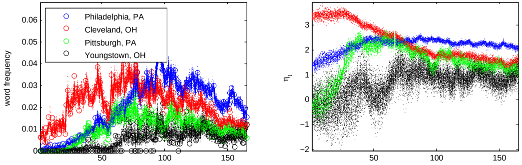
$\eta_3 \sim N(A\eta_2, \Sigma)$

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

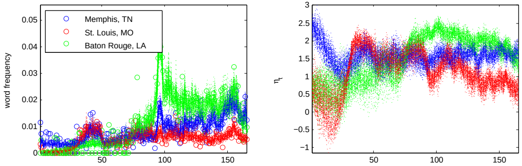
$$\begin{aligned} c_{\text{ctfu},2} &\sim \text{Binomial}(f(\eta_{\text{ctfu},2}), N_2) \\ c_{\text{hella},2} &\sim \text{Binomial}(f(\eta_{\text{hella},2}), N_2) \\ &\ldots \end{aligned}$$

Estimating parameters of this autoregressive process reveals geographic pathways of diffusion across thousands of words (Eisenstein et al., 2014).

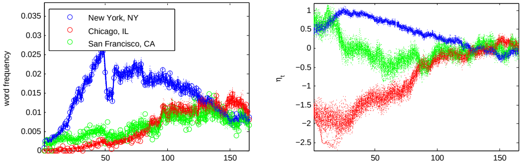
ctfu


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

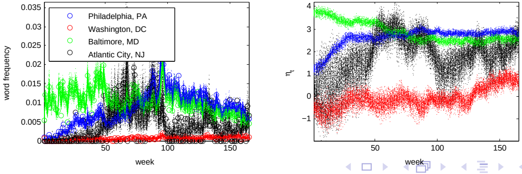
ion


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

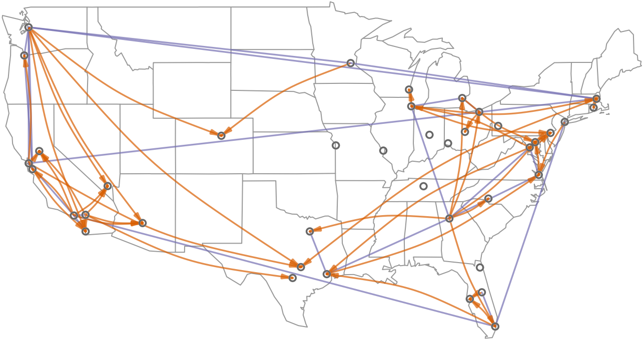
-_-


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

ard


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

Aggregating region-to-region influence



Highly-confident pathways of diffusion (from autoregressive parameter A).

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.

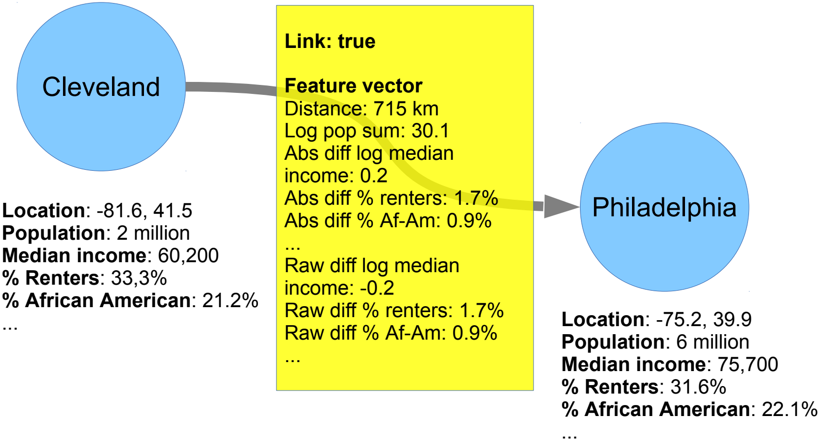
- 2010 US Census gives detailed demographics for each city.
- Are there types of demographic relationships that are especially frequent among linked cities?

Logistic regression

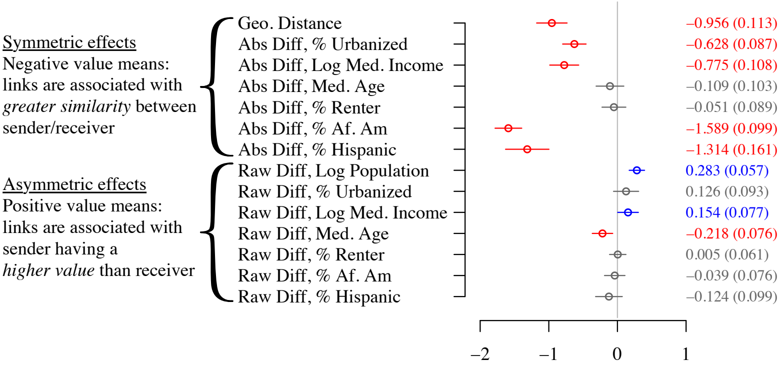
Location: -81.6, 41.5
Population: 2 million
Median income: 60,200
% Renters: 33,3%
% African American: 21.2%
...

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

Logistic regression



Regression coefficients



- Assortativity by race (of cities!) even more important than geography.
- Asymmetric effects are weaker, but bigger, younger metros tend to lead.

References I

Adamic, L. A. & Adar, E. (2003). Friends and neighbors on the web. *Social networks*, *25*(3), 211–230.

Backstrom, L., Sun, E., & Marlow, C. (2010). Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of WWW*, (pp. 61–70). ACM.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.

Blundell, C., Beck, J., & Heller, K. A. (2012). Modelling reciprocating relationships with hawkes processes. In F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25* (pp. 2600–2608). Curran Associates, Inc.

Bourdieu, P. (1984). *Distinction: A social critique of the judgement of taste*. Harvard University Press.

Crystal, D. (2006). *Language and the Internet* (Second ed.). Cambridge University Press.

Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics*, *12*(4), 453–476.

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2010). A latent variable model for geographic lexical variation. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, (pp. 1277–1287)., Stroudsburg, Pennsylvania. Association for Computational Linguistics.

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

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, *58*(1), 83–90.

Huberman, B., Romero, D. M., & Wu, F. (2008). Social networks that matter: Twitter under the microscope. *First Monday*, *14*(1).

Pavalanathan, U. & Eisenstein, J. (2015). Audience-modulated variation in online social media. *American Speech*, *90*(2).

Puniyani, K., Eisenstein, J., Cohen, S., & Xing, E. P. (2010). Social links from latent topics in microblogs. In *Proceedings of NAACL Workshop on Social Media*, Los Angeles.

Trudgill, P. (1972). Sex, covert prestige and linguistic change in the urban british english of norwich. *Language in Society*, *1*(2), 179–195.