Socio-Digital Influence Networks from Language Analysis
(FA9550-14-1-0379)

PI: Jacob Eisenstein

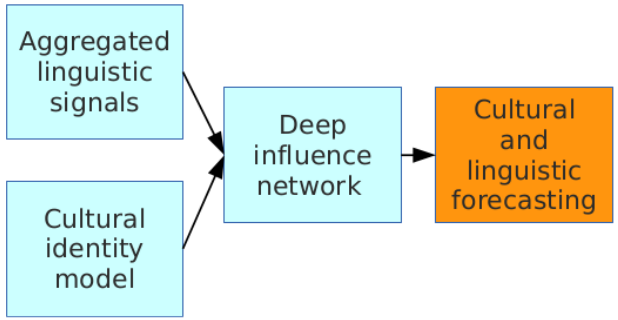
(Georgia Institute of Technology;

School of Interactive Computing)

AFOSR Program Review: Trust & Influence
June 13-17, 2016

Project Summary

Research Objectives:



Induce actionable patterns of social power and influence from socio-digital information traces, forecasting future changes in language and culture..

Technical Approach:

- Mine cultural artifacts and socio-digital information sources for linguistic signals and social behavioral traces.
- Invent machine learning techniques for **structure induction** to find latent structures that concisely explain both modalities.
- Apply longitudinal and causal analysis to identify how sociocultural influence spreads over these structures, yielding actionable predictions.

Key Findings:

- We can quantify latent social influence from timestamped social media traces by using point process models.
- Influence detection can be scaled to networks of millions of individuals by identifying relevant social covariates, and estimating their predictiveness.
- Social network tie strength, as measured by embeddedness, predicts linguistic influence
- The role of geographical proximity is less clear; geographical assortativity may fully explain regional linguistic differences.

Benefits to the wider academic or DoD community:

1. New mathematical models for determining the structural factors underlying sociocultural influence.
2. New algorithms for performing efficient estimation in these models.
3. Substantive insights on how social influence drives processes of language change.

Project Start Date: October 2014

Project End Date: September 2017

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

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

Language change as a sociocultural process

Traces of language change reveal the structure and character of latent influence networks.

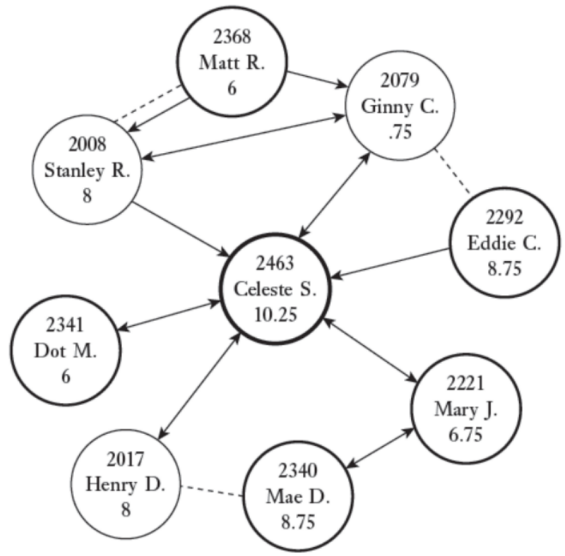
- **Exposure**: To use a new word, you must be exposed to it.
- **Influence**: The *choice* of whether to use a new word is a socially-motivated decision.

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 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

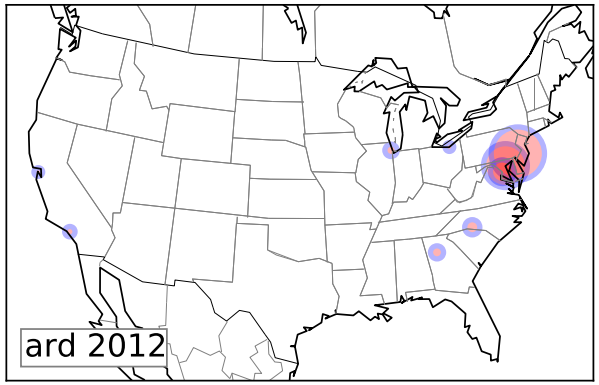
We focus on nine metropolitan areas in the USA:

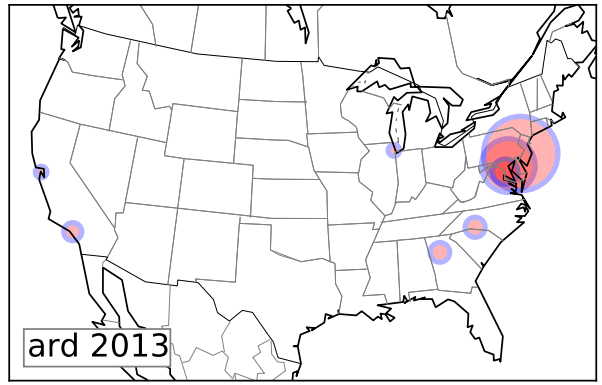
- Atlanta
- Baltimore
- Charlotte
- Chicago
- Cleveland
- Los Angeles
- Philadelphia
- San Francisco
- Washington, DC

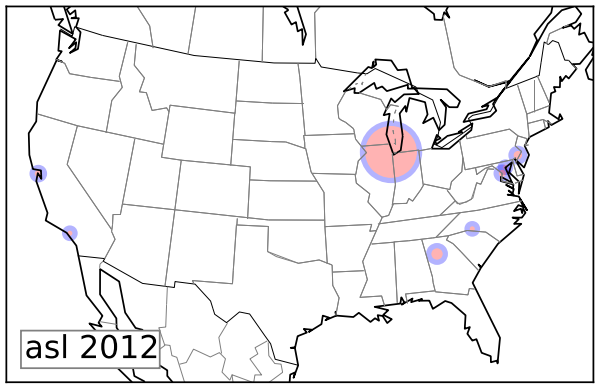
Linguistic variables

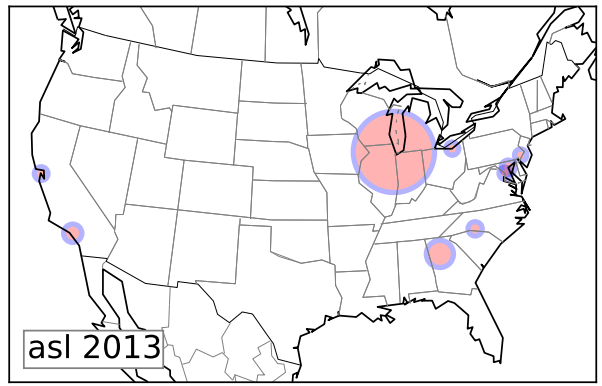
	count	type	definition
ard	28,882	phonetic	alternative spelling for alright, e.g., lol ard ill text u
asl	36,159	abbreviation	intensifier, e.g., I'm hungry asl
dese	1,664	phonetic	alternative spelling for these, e.g., I ain't like sum of dese frauds
hella	20,470	lexical	intensifier, e.g., I'm hella hungry
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

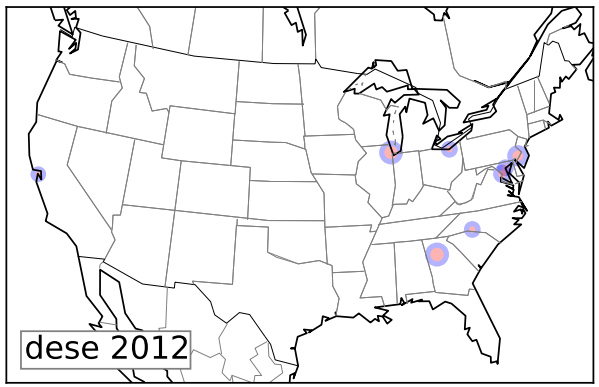
Linguistic variables

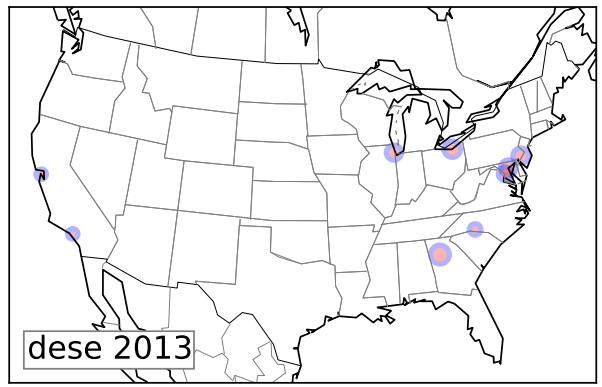




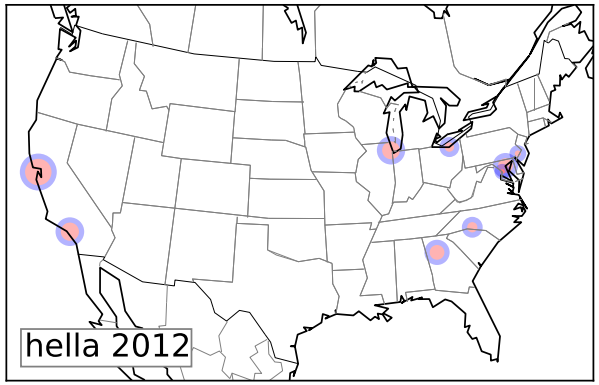


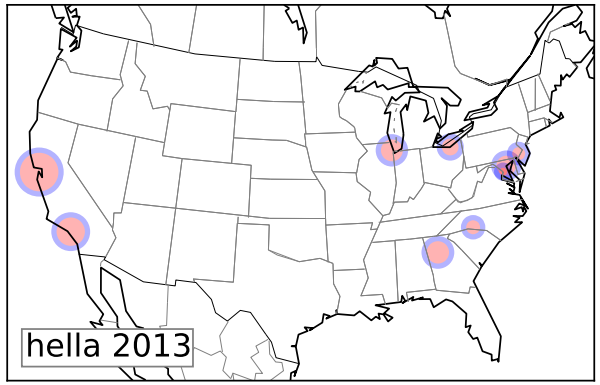


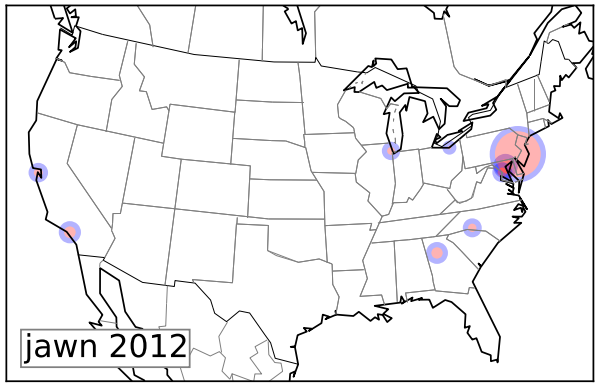


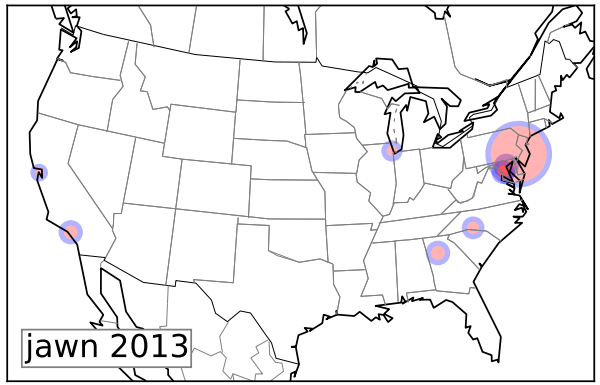


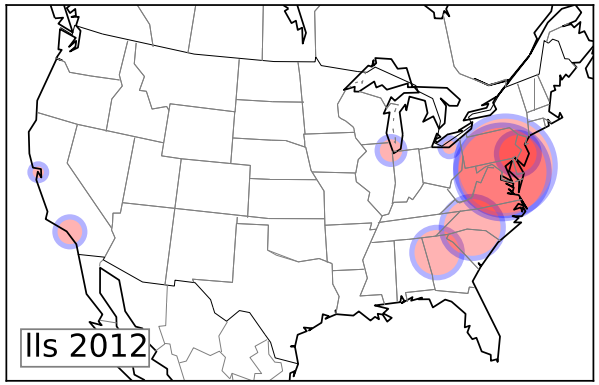
Linguistic variables

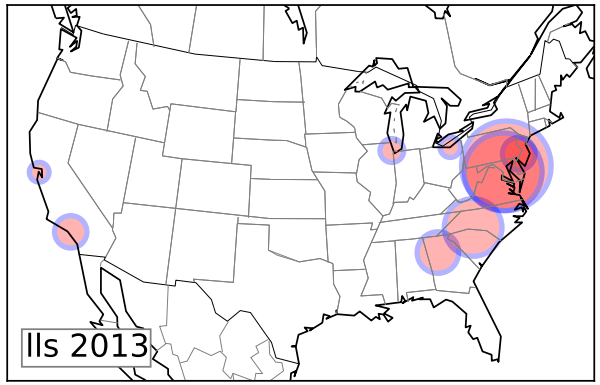












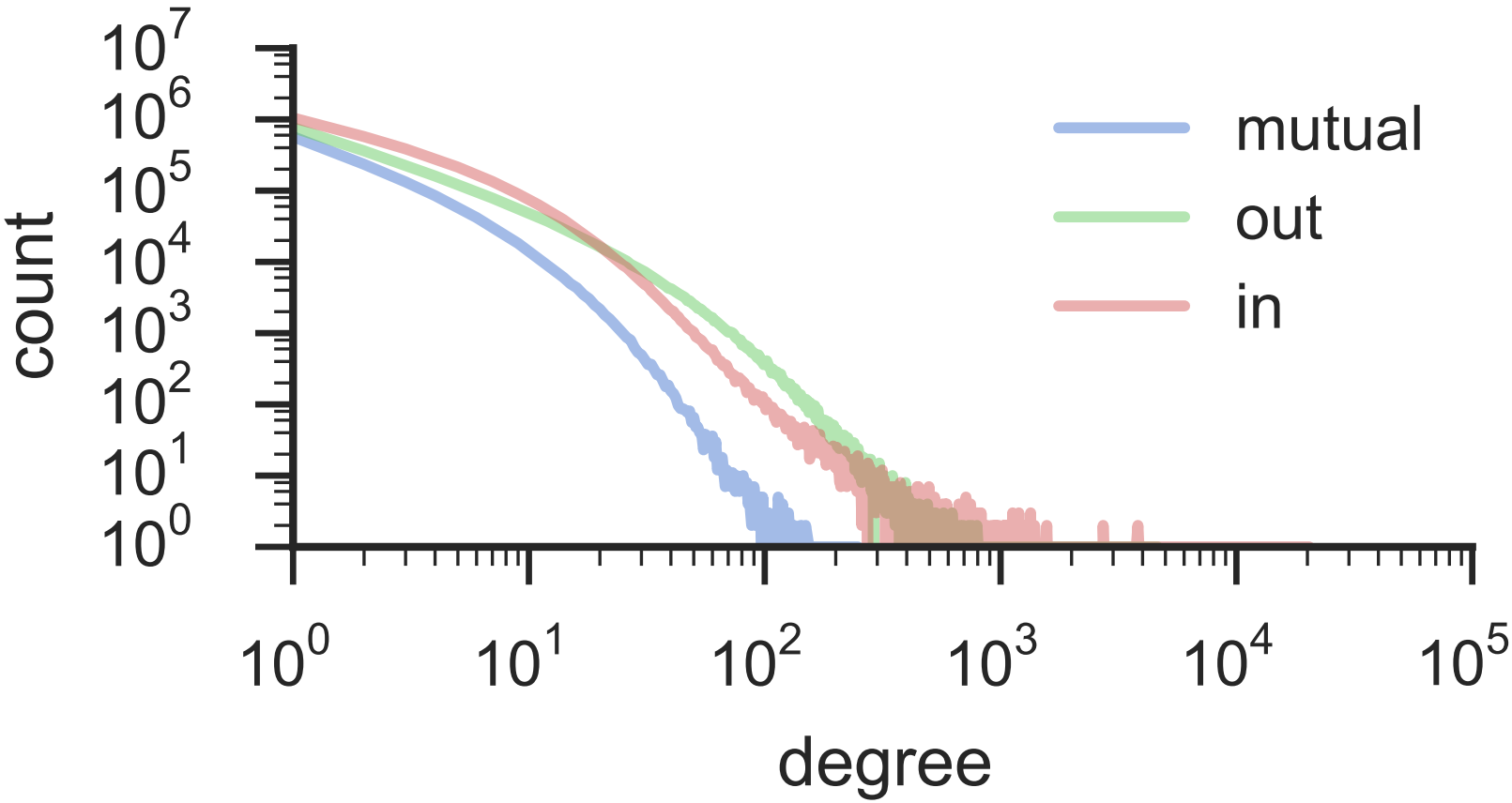
Social network

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 are better conduits of language change.

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

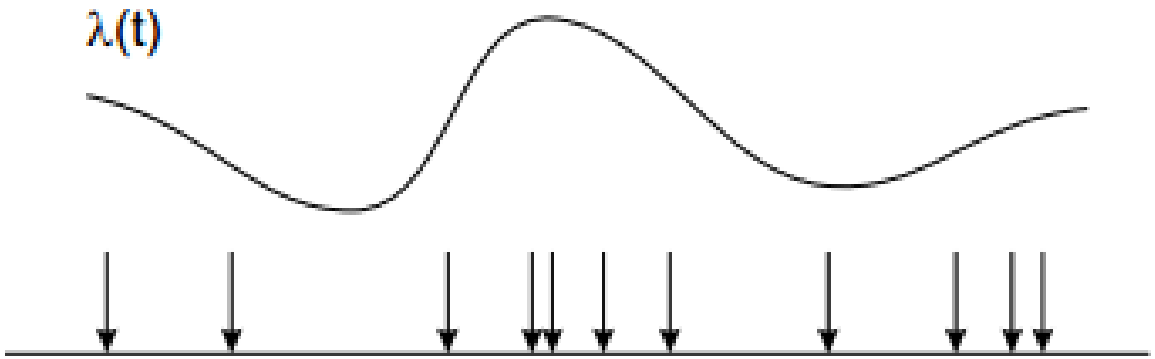
The Poisson process

- 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)

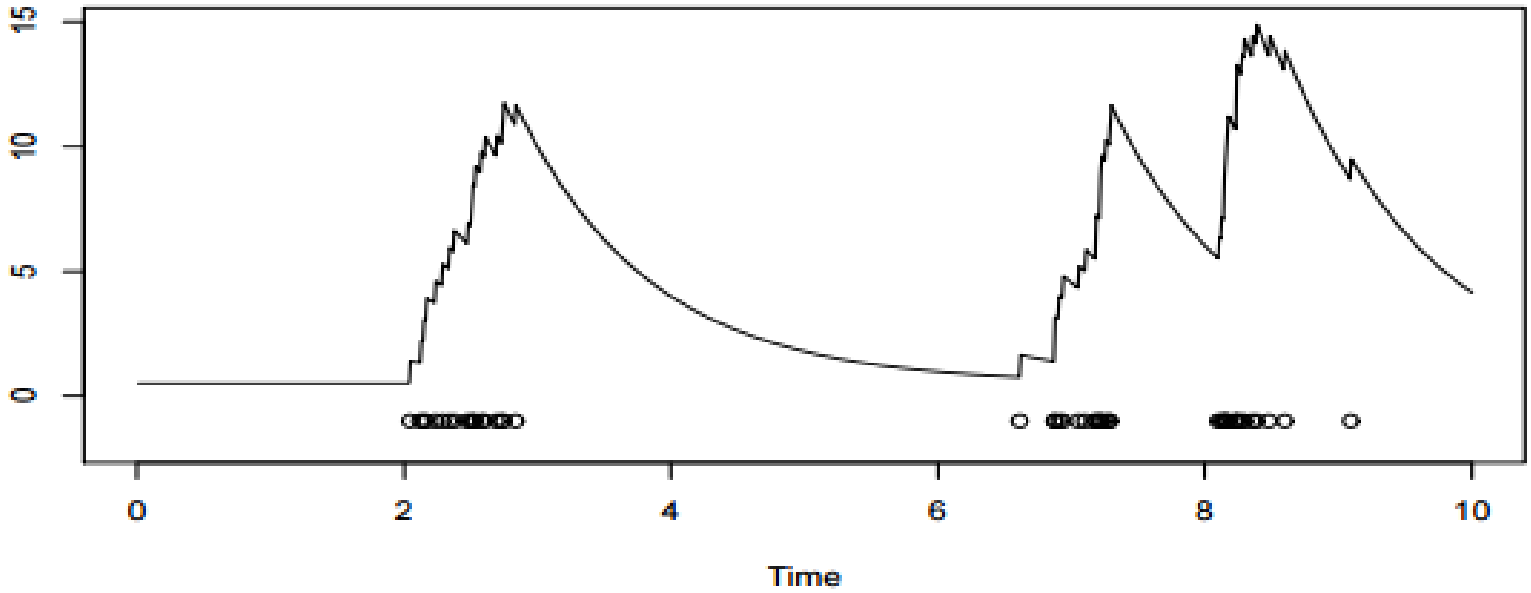
$$\lambda(t) = \mu + \alpha \sum_{t_n < t} \kappa(t - t_n), \tag{3}$$

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

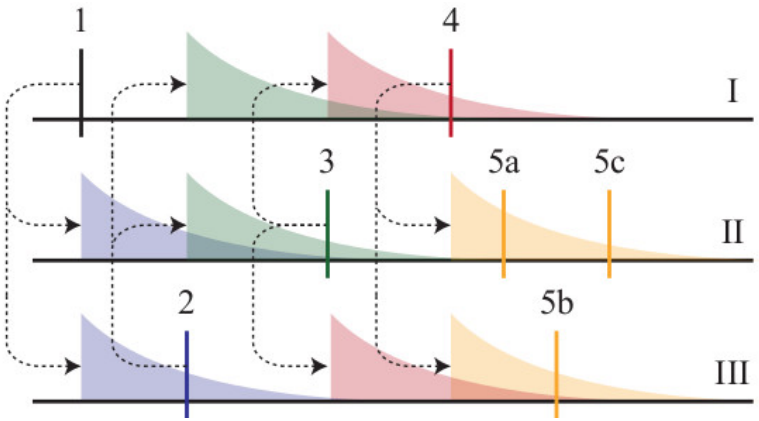
$$\lambda_m(t) = \mu_m + \sum_{t_n < t} \alpha_{m_n \to m} \kappa(t - t_n), \qquad (5)$$

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

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

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

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

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

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

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.

Measuring tie strength

Mutual friends

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

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. Linguistic influence passes between 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.

Some next steps

- Control for exogenous factors, using search query records.
- Use parametric Hawkes Process to quantitatively compare procedures for social network construction and metrics for tie strength.
- Generalizing from linguistic influence: transmission of hate speech.

Thanks

To the **AFOSR Trust and Influence program** and to my co-authors and collaborators,

- **Georgia Tech**: Rahul Goel, Sandeep Soni, Naman Goyal, Le Song;
- **Columbia**: John Paparrizos;
- **Microsoft Research**: Hanna Wallach, Fernando Diaz;
- **ENS-Lyon**: Márton Karsai.

References I

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

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.

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).

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.

Publications, Awards, Patents, or Transitions Attributed to the Grant

- **A Parametric Hawkes Process Model of Language Change in Online Social Networks**. Rahul Goel, Sandeep Soni, Naman Goyal, John Paparrizos, Hanna Wallach, Fernando Diaz, and Jacob Eisenstein (in prep)
- **Putting Things in Context: Community-specific Embedding Projections for Sentiment Analysis**. Yi Yang and Jacob Eisenstein (in review)
- **Toward Socially-Infused Information Extraction: Embedding Authors, Mentions, and Entities.** Yi Yang and Jacob Eisenstein (in review)
- **Confounds and Consequences in Geotagged Twitter Data**. Umashanthi Pavalanathan and Jacob Eisenstein. Proceedings of Empirical Methods in Natural Language Processing (EMNLP), November 2015.
- Invited presentations/keynotes
  - Workshop on Language and Social Networks, at École Normale Supérieur, Lyon, May 2016.
  - Text as Data Colloquium series at NYU, March 2016.
  - Computational Social Science Colloquium, Michigan State University, February 2016.
  - IGERT Distinguished Speaker Series, Columbia University, November 2015.