Social Dynamics of Language Change in Online Networks

Sandeep Soni [1] Rahul Goel [1] Naman Goyal [1]
John Paparrizos [2]
Hanna Wallach [3] Fernando Diaz [3] Jacob Eisenstein [1]

[1]Georgia Institute of Technology

[2]Columbia University

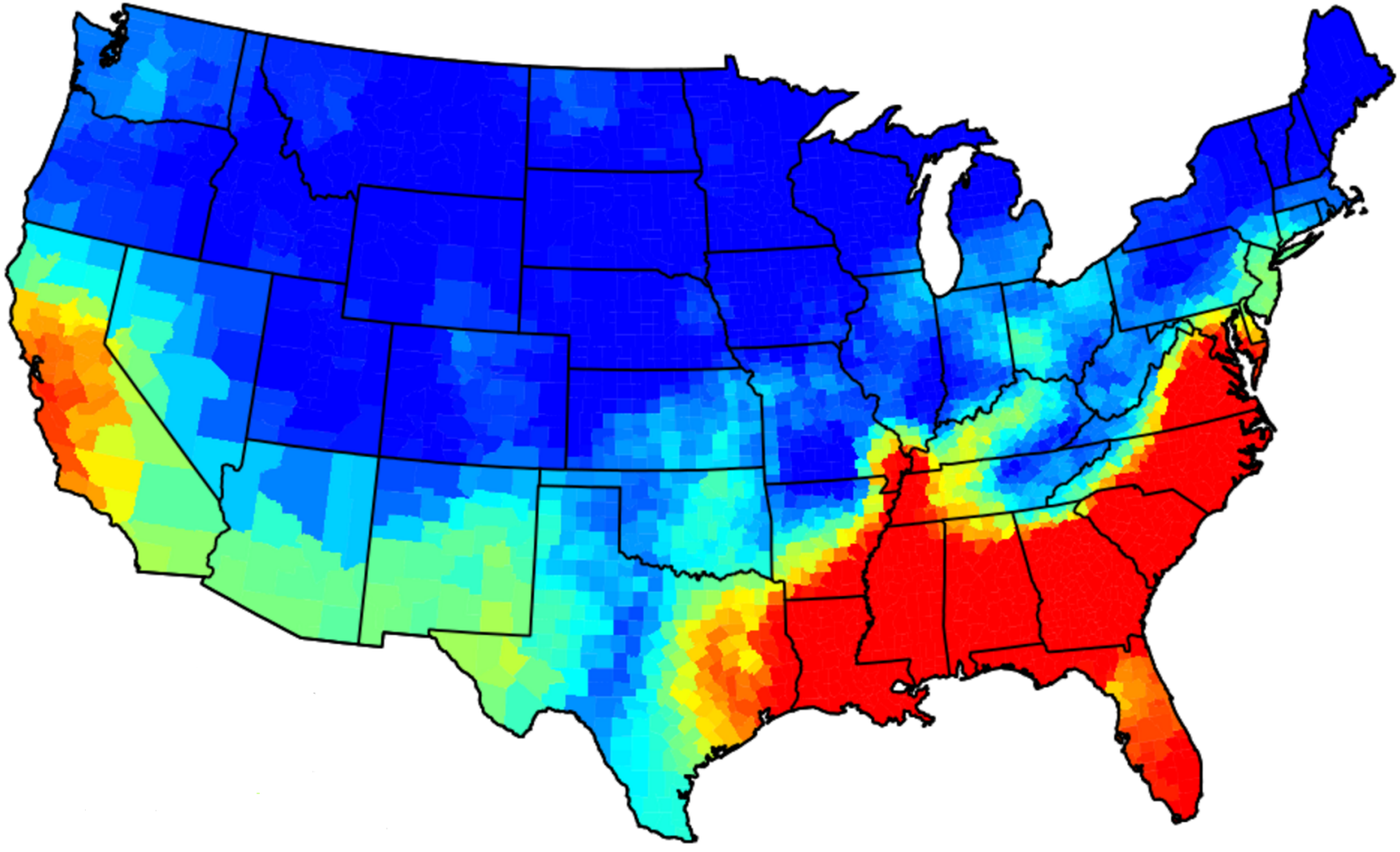
[3]Microsoft Research

November 17, 2016

Dwayne Johnson
@TheRock

Follow

Doin' some sanging over here or as we say in the streets "I'm spittin' that fire bruh!"🎤🔥😉 w/ @Lin_Manuel. #YoureWelcome #MoanaMondays 🌊🎼



What does it take for someone to use a non-standard word or neologism, like bruh?

Language change is a social phenomenon

For a newer linguistic form to spread successfully, two things must happen:

Language change is a social phenomenon

For a newer linguistic form to spread successfully, two things must happen:

- **Exposure**: you must have come in contact with someone using it before.
- **Adoption**: you must have the willingness to adopt the linguistic form.

Language change is a social phenomenon

For a newer linguistic form to spread successfully, two things must happen:

- **Exposure**: you must have come in contact with someone using it before.
- **Adoption**: you must have the willingness to adopt the linguistic form.

Studying language change can reveal the social structure: communication and influence pathways in a social network.

Language change and social networks

Close-knit networks, which vary in the extent to which they approximate to an idealized maximally dense and multiplex network, have the capacity to maintain and even enforce local conventions and norms - including linguistic norms - and can provide a means of opposing dominant institutional values and standardized linguistic norms.

- J.Milroy and L.Milroy

The leaders of linguistic change are people at the center of their social networks, who other people frequently refer to, with a wider range of social connections than others.

- William Labov

Hypotheses

This work uses large scale Twitter data to test the following hypotheses:

- **H1:** Language change spreads through online social network connections.
- **H2:** Densely embedded ties are more linguistically influential.
- **H3:** Geographically local ties are more linguistically influential.

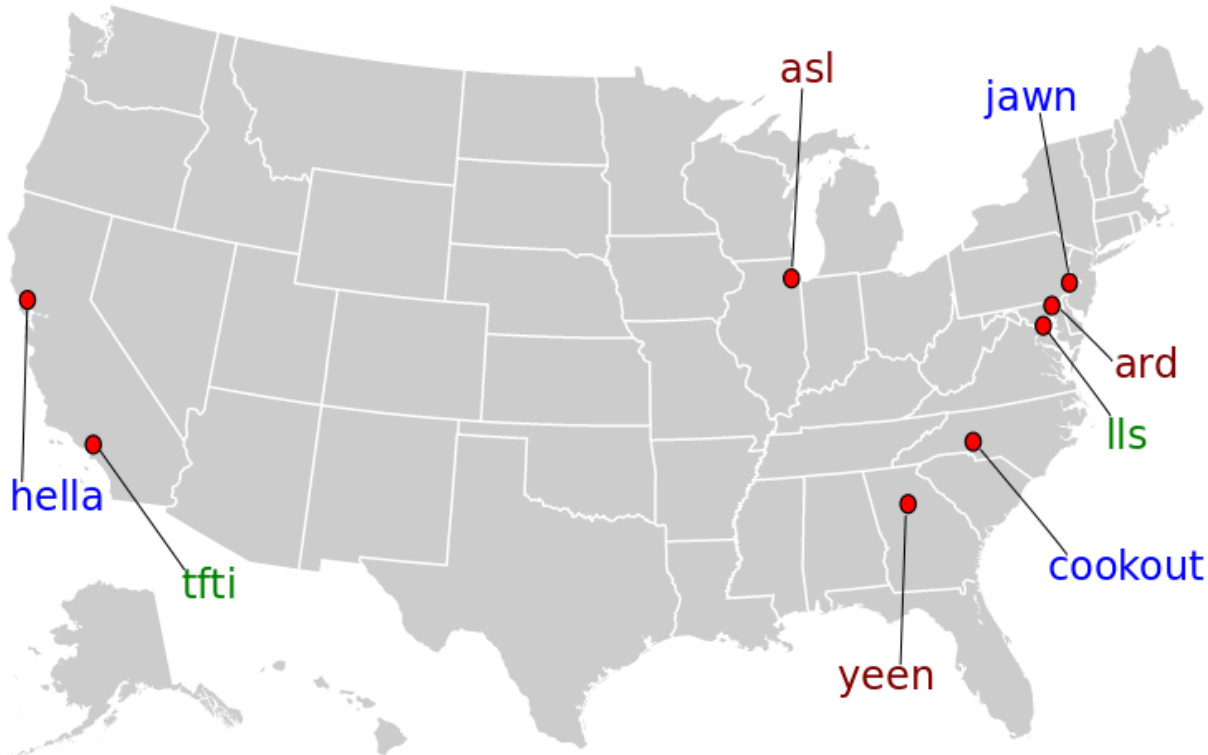
Data

We use a large scale Twitter dataset

- **Period**: All public messages from US between June 2013 and June 2014.
- **Scale**: A total of 4.35 million twitter users with geolocation and social network metadata.
- **Linguistic variables**: 16 non-standard words divided into 3 linguistic categories.

Data: linguistic variables and geography

- These words are known to be strongly associated with the selected areas.



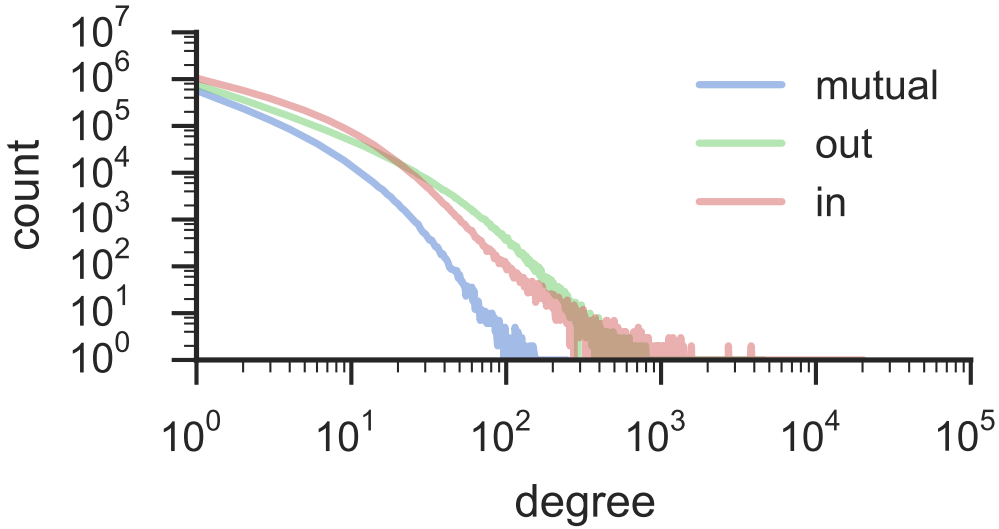
For example:

- ard : phonetic spelling for all right.
- hella: lexical word typically means extremely.
- lls: abbreviation for laughing like shit.

Data: social network

A social network connection exists between two users if they have mentioned each other.

- sarah: @mark going to the game?
- mark: @sarah yea



The mentions network is more easily available and is socially meaningful than follower-followee network

Dataset summary

Social network

Mark	Sarah
Mark	Mickey
Sarah	Todd
Todd	Barney
...	...

Locations

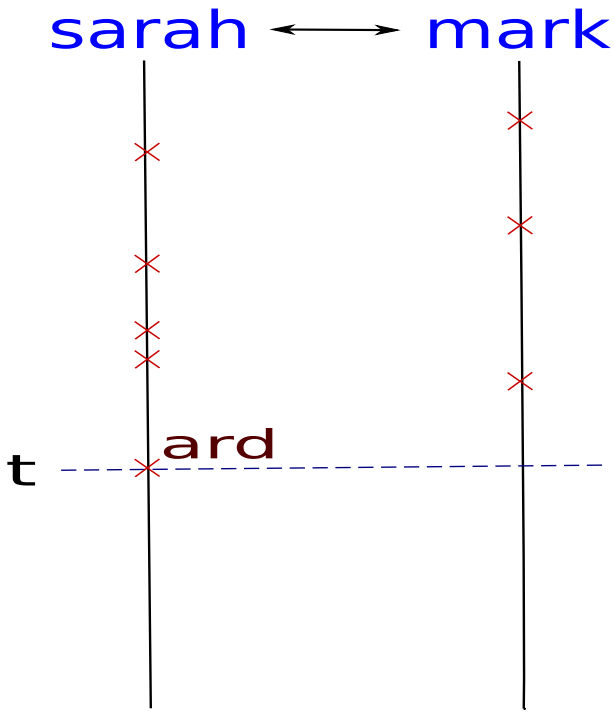
Mark	Los Angeles
Mickey	Los Angeles
Sarah	Atlanta
Todd	Chicago
...	...

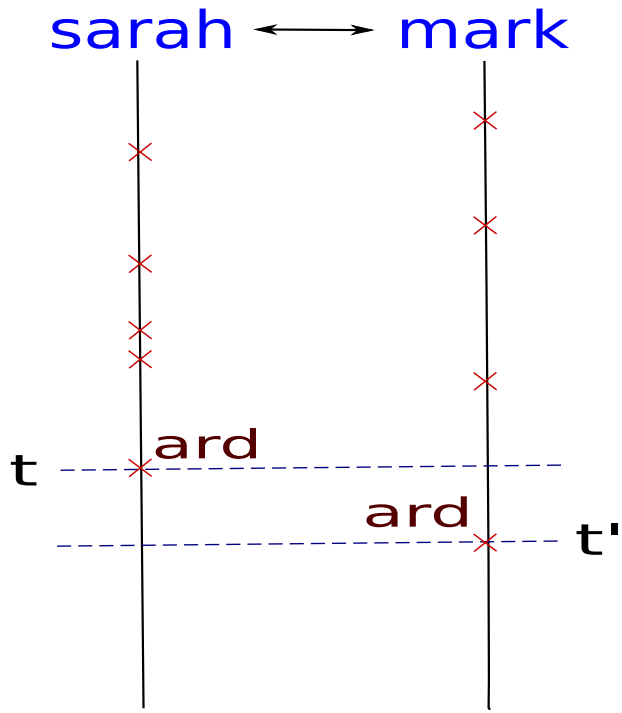
Language

Mark	jawn	Jun 1, 2013, 13:45
Mickey	jawn	Jun 1, 2013, 13:50
Todd	hella	Jun 1, 2013, 18:15
Mark	lls	Jun 2, 2013, 07:30
Mickey	lls	Jun 2, 2013, 07:40
...	...	...

Language change as a contagion

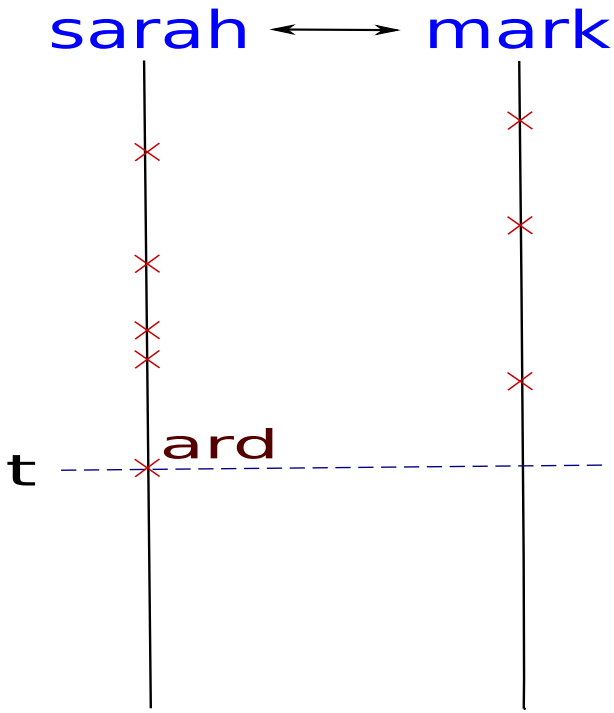
H1: Language change spreads through online social network connections.

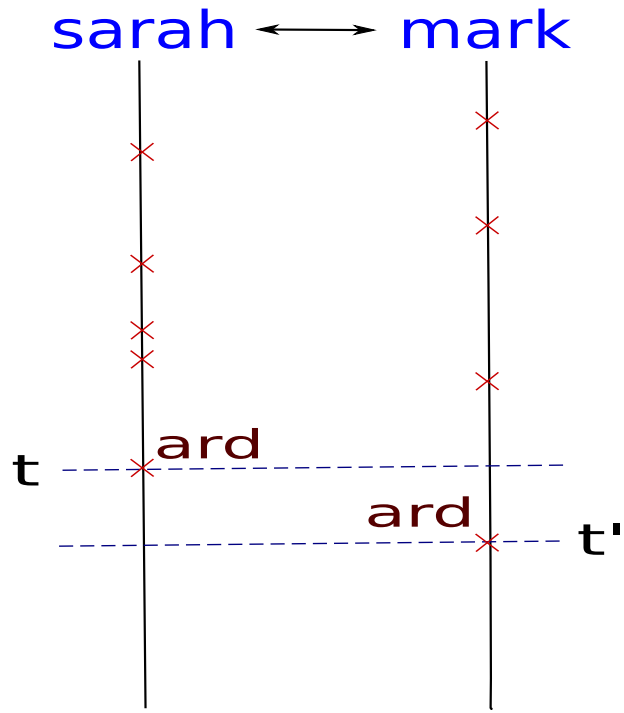




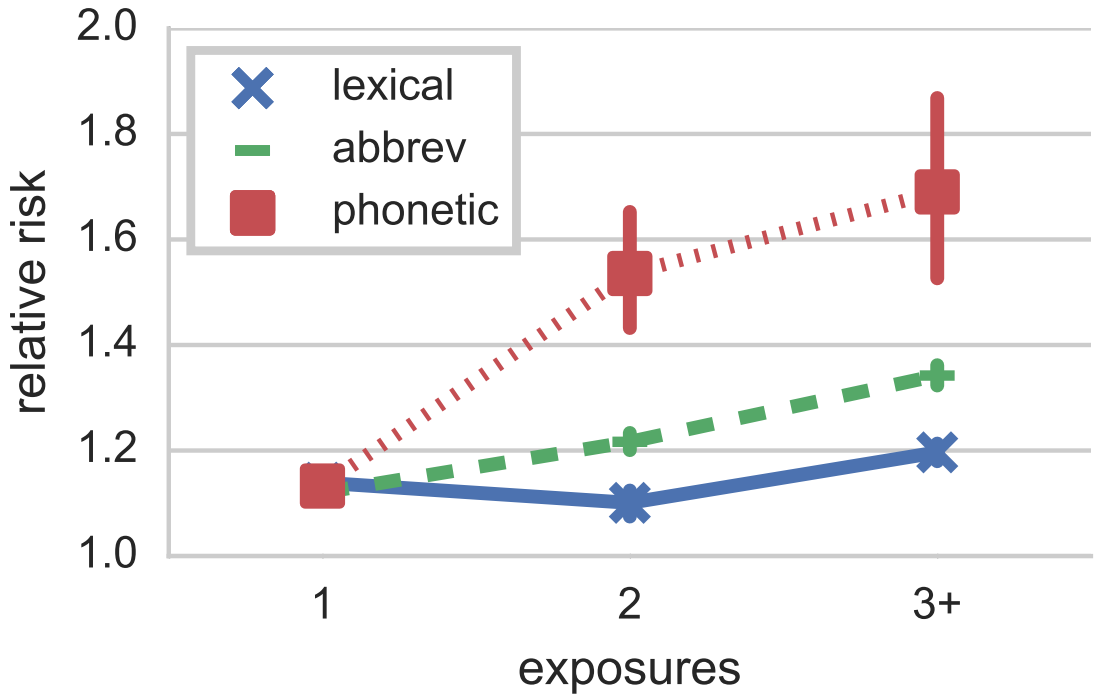
Language change as a contagion

H1: Language change spreads through online social network connections.





$$\text{Infection risk (w)} = \frac{\text{\# adopters after exposure}}{\text{\# users exposed}}$$



- Risk ratios for all the three classes is greater than 1 implying social influence at play in the spread of language change online.

- The previous analysis reveals linguistic influence through social network ties.

  But! we did not account for:

  - Differences in the types of network ties (Hypothesis 2 and 3).
  - Differences in exposure times: recent exposure may be more influential.
  - Differences in users' activity rates: some users just tweet more than others.

- The previous analysis reveals linguistic influence through social network ties.

  But! we did not account for:

  - Differences in the types of network ties (Hypothesis 2 and 3).
  - Differences in exposure times: recent exposure may be more influential.
  - Differences in users' activity rates: some users just tweet more than others.

- We use statistical models that incorporate the above ideas and we evaluate them by how good they fit the data.

Multivariate Hawkes process

- For every word, a cascade of events about when the word was used and who used it is $\{t_n, m_n\}_{n \in 1 \ldots N}$.
- Every user has a intensity function to define how likely it is to use a word at any time.

$$\lambda_m(t) = \mu_m + \sum_{t \leq t_n} \alpha_{m_n \to m} \kappa(t - t_n), \tag{1}$$

where $\kappa(\Delta t)$ is usually defined as,

$$\kappa(\Delta t) = \exp^{-\gamma \Delta t} \tag{2}$$

- μ_m is the base rate of user m
- $\alpha_{m_n \to m}$ is excitation from events by user m_n on m.
- γ is the time scale.

Parametric Hawkes process

- **But!** number of parameters is quadratic in the number of users.
- We make α to be a function of shared features between each pair of individuals.

$$\alpha_{m_1 \to m_2} = \theta^T \mathbf{f}(m_1 \to m_2) \tag{3}$$

- Now we only need to estimate $\#|\theta|$ parameters instead of M^2

Features

Self-excitation $f_1(m_1, m_2) = 1$ if $m_1 = m_2$.

Mutual friend $f_2(m_1, m_2) = 1$ if $m_1 \leftrightarrow m_2$.

Tie strength $f_3(m_1, m_2) = 1$ if m_1 and m_2 are mutual friends **and** have a *strong* tie.

- Strength of the tie is measured by the Adamic-Adar measure.
- A tie between a pair is strong if its strength is greater than 90 % other pairs.

Locality $f_4(m_1, m_2) = 1$ if m_1 and m_2 are mutual friends **and** geolocated to the same metropolitan area.

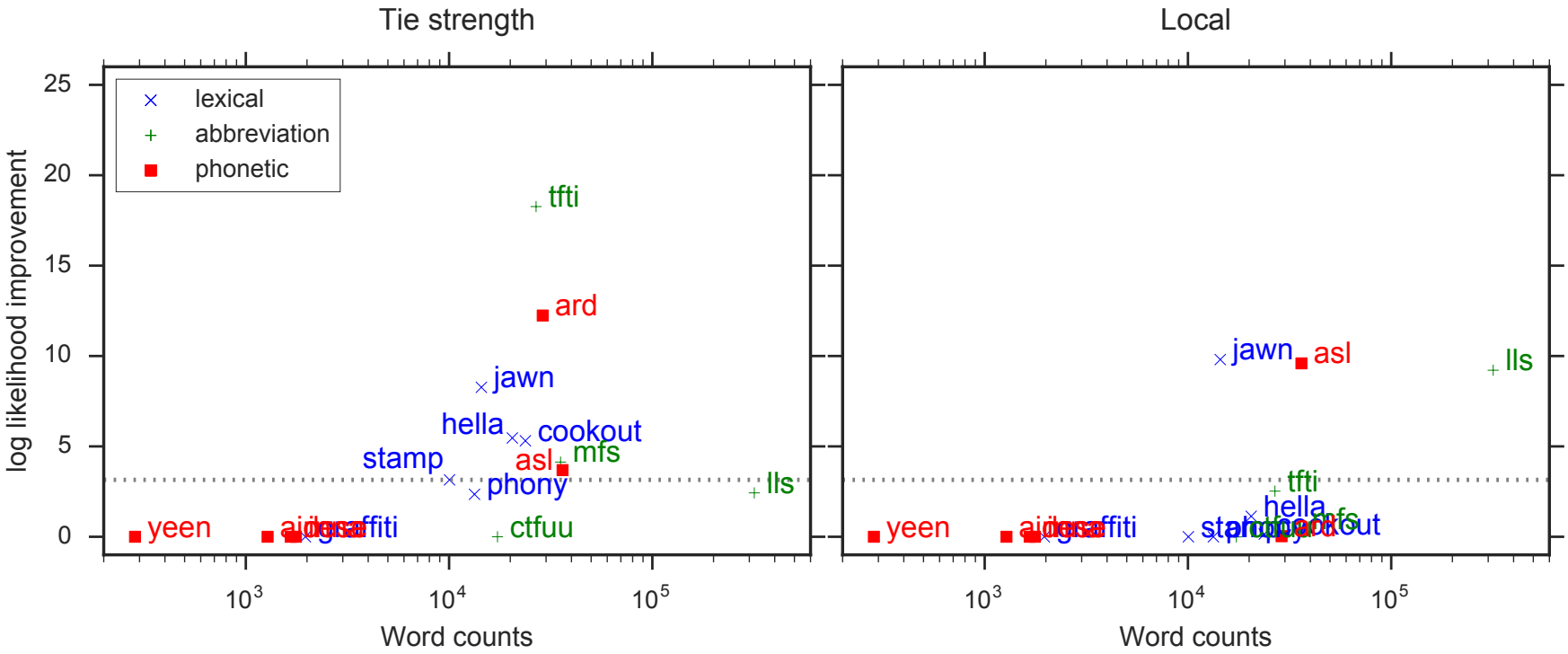
Hypothesis testing

We create multiple nested models:

- Null model: F1 + F2
- Null Vs F1 + F2 + F3 : are densely embedded ties more linguistically influential?
- Null Vs F1 + F2 + F4 : are geographically local ties more linguistically influential?

  We compare models using a likelihood ratio test, with correction for multiple comparisons.

Results



- Linguistic influence exerted across densely embedded ties greater than linguistic influence exerted across other ties.
- But little evidence to suggest that linguistic influence exerted across geographically local ties.

Discussion and future work

- **Social network matters** : Users are likely to use a new word if their friends have used it before.
- **Strong ties matter more** : Close friends have more linguistic influence.
- **Shared geography matters less** : Weak evidence that Twitter users pay special attention to geographically local ties.

What next?

- Is linguistic influence correlated with other forms of influence?
- Is the diffusion process of language change different from other forms like hashtags or URLs?

hella thanks *yall*!