Variation and Change in Online Social Networks

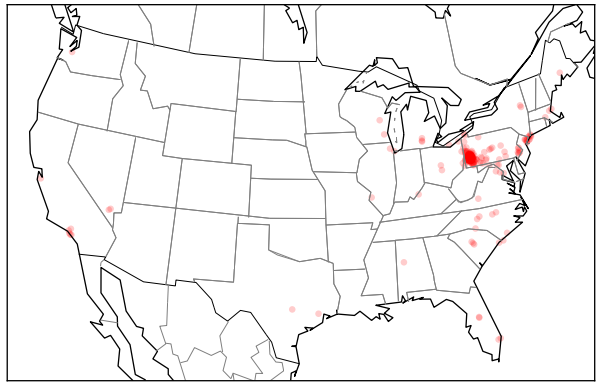
Jacob Eisenstein
@jacobeisenstein

Georgia Institute of Technology

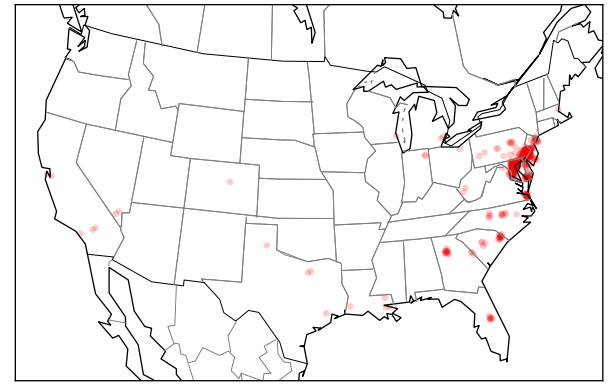
April 21, 2017

The important of place in online writing

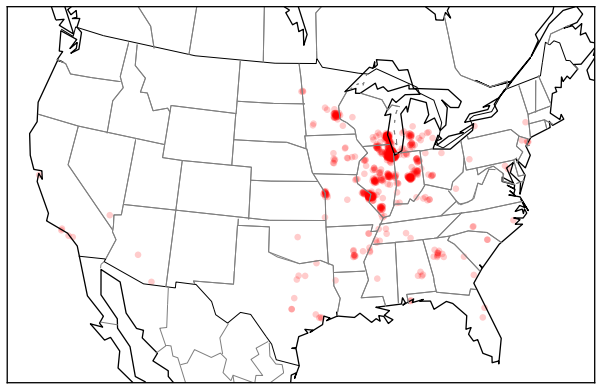
yinz



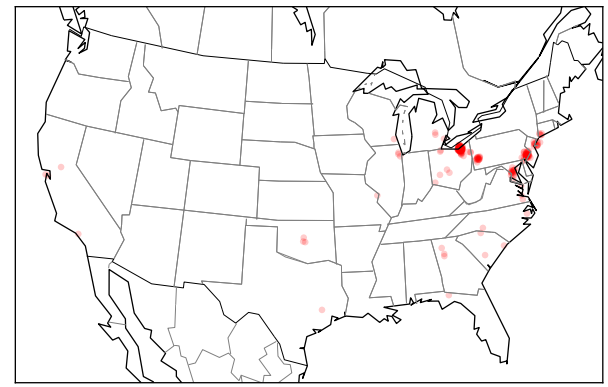
ard ("alright")



lbvs ("laughing but very serious")

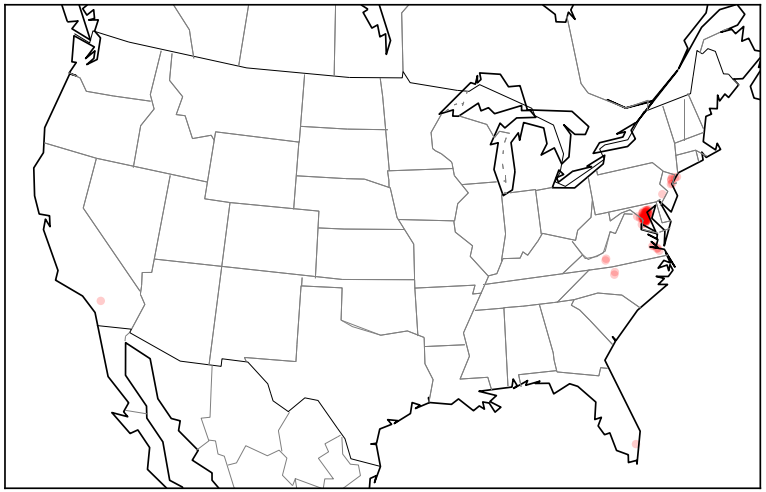


ctfu ("cracking the fuck up")

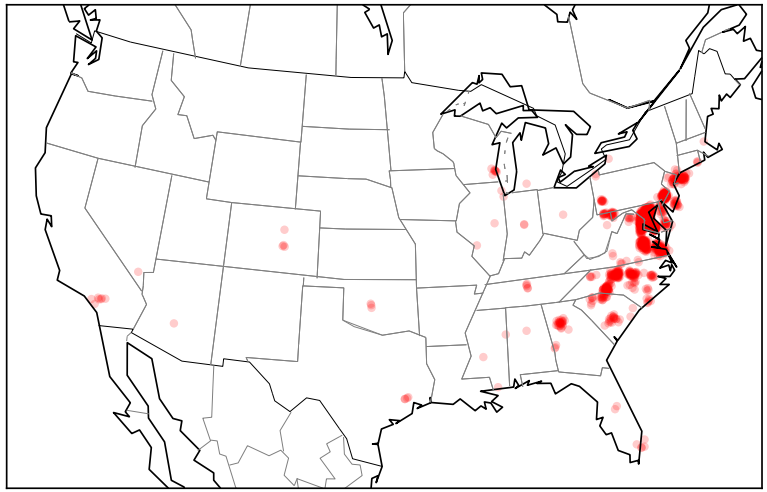


(Eisenstein et al., 2010, 2014)

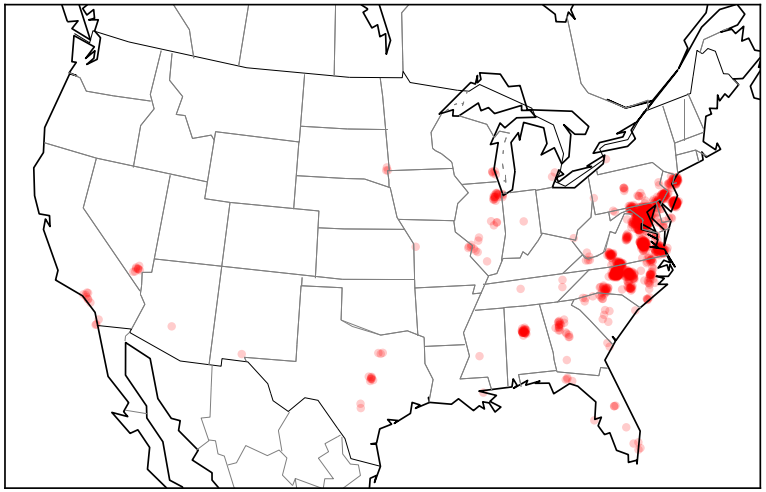
DC's contribution: lls



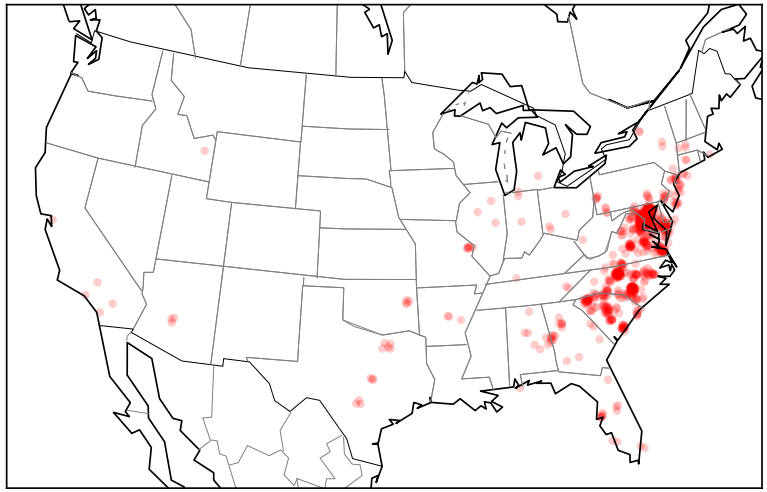
2009



2010



2011

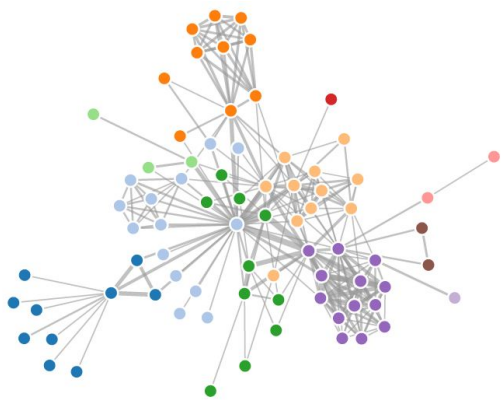


2012

Language change in the network

How does these macro-scale patterns of variation and change ground out in individual social networks?

- Can macro-scale phenomena be explained in terms of individual social choices?
- How does network structure affect the trajectory of language change?



Hypotheses

Goel et al. (2016): use large-scale Twitter data to test three hypotheses about language change.

- **H1**: language change is transmitted across social media networks.
- **H2**: strong ties are better conduits of language change.
- **H3**: geographically local social network ties are better conduits of language change (*covert prestige*; Trudgill, 1972; Bourdieu, 1984).

Dataset

- Twitter analysis is usually conducted on a **sample** from the streaming API (e.g., Eisenstein et al., 2010; Huang et al., 2016).
- Modeling the fine structure of language change requires **complete data**, because random samples 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.

Cities and distinctive features

- Atlanta: ain, dese, yeen
- Baltimore: ard, inna, lls, phony
- Charlotte: cookout
- Chicago: asl, mfs
- Los Angeles: graffiti, tfti
- Philadelphia: ard, ctfuu, jawn
- San Francisco: hella
- Washington, DC: inna, lls, stamp

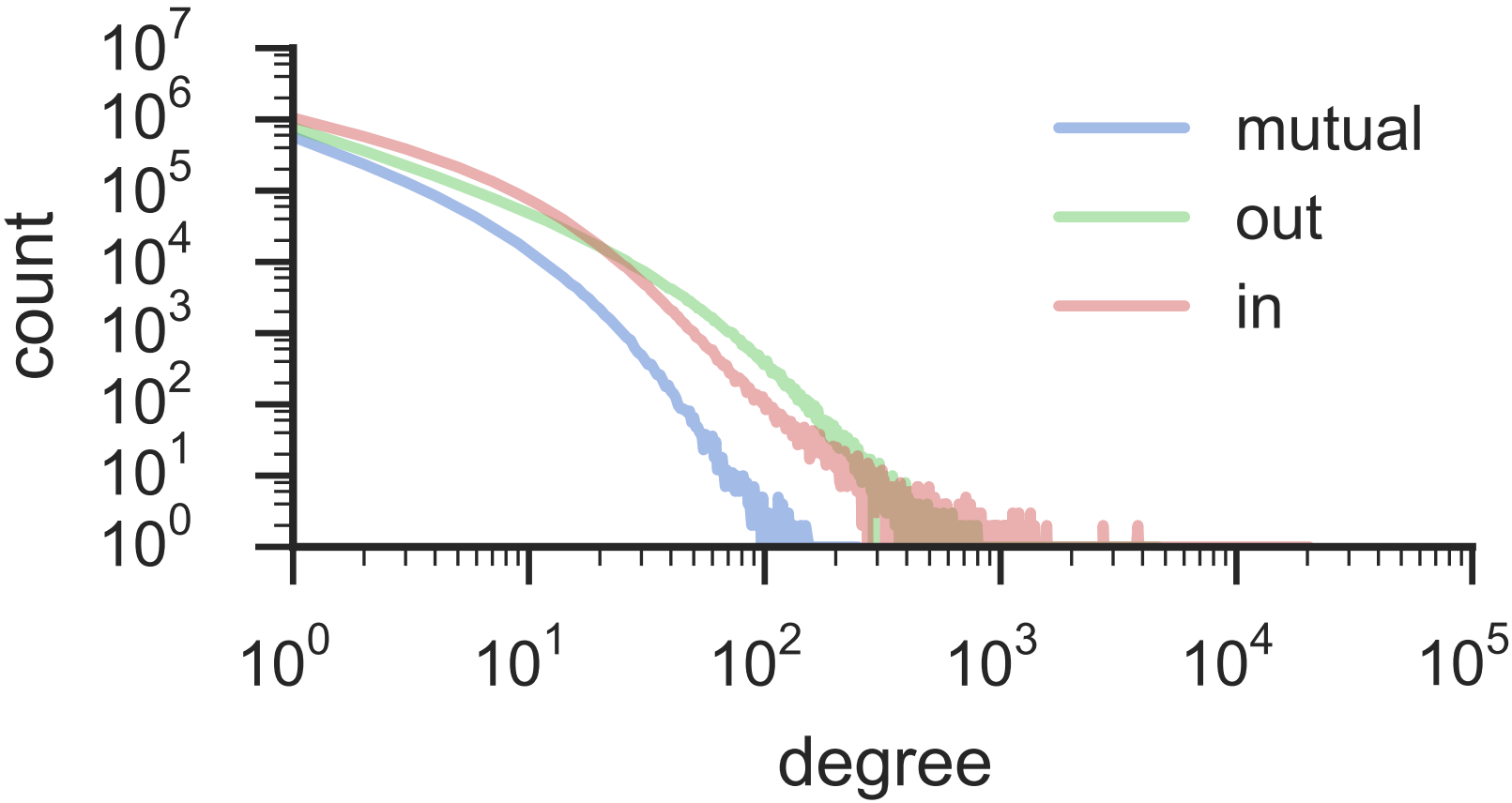
Social network

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* is more socially meaningful than the *articulated network* of follower-followee links (Huberman et al., 2008).

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
...	...	...

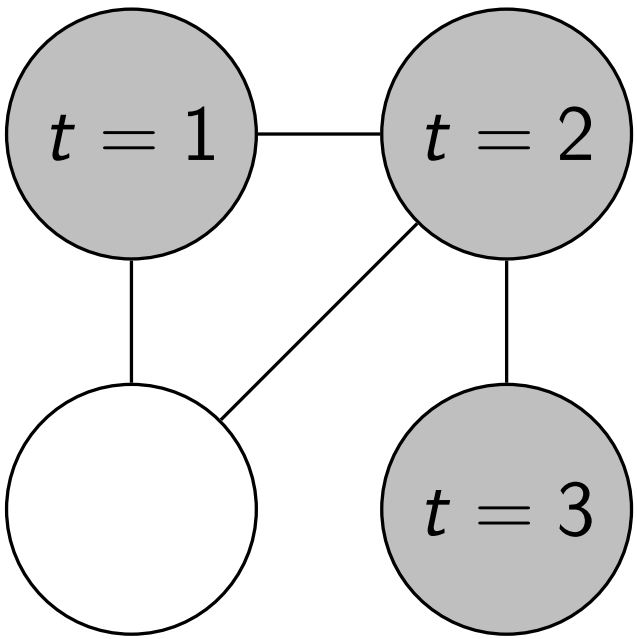
Hypotheses

- **H1**: Language change is transmitted across social media networks.
- **H2**: Strong ties are better conduits of language change.
- **H3**: Geographically local social network ties are better conduits of language change.

Hypotheses

- **H1**: Language change is transmitted across social media networks.
- **H2**: Strong ties are better conduits of language change.
- **H3**: Geographically local social network ties are better conduits of language change.

Shuffle test for influence

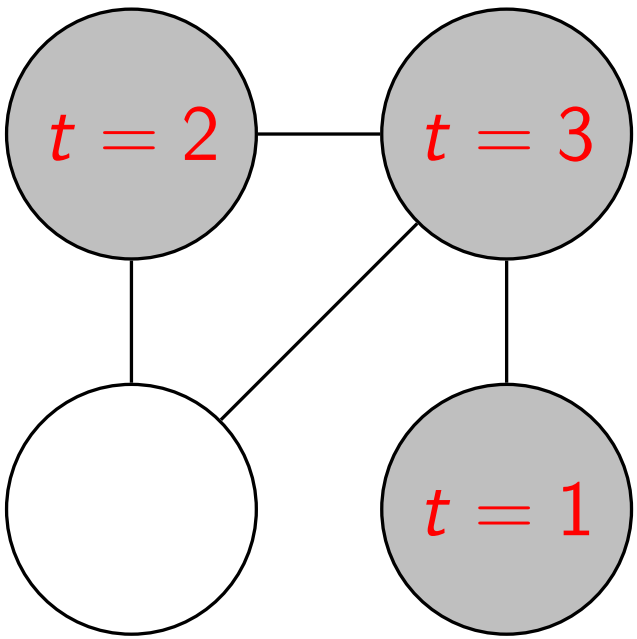


- Observed data

$$P(\text{infection} \mid 0 \text{ exposures}) = \frac{1}{4}$$

$$P(\text{infection} \mid \geq 1 \text{ exposures}) = \frac{2}{3}$$

Shuffle test for influence



- Observed data

$$P(\text{infection} \mid 0 \text{ exposures}) = \frac{1}{4}$$

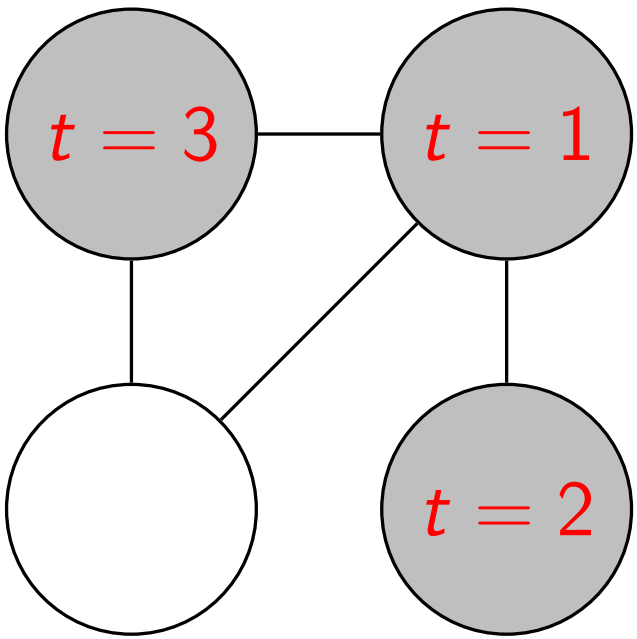
$$P(\text{infection} \mid \geq 1 \text{ exposures}) = \frac{2}{3}$$

- Randomized data

$$P(\text{infection} \mid 0 \text{ exposures}) = \frac{2}{4}$$

$$P(\text{infection} \mid \geq 1 \text{ exposures}) = \frac{1}{2}$$

Shuffle test for influence



- Observed data

$$P(\text{infection} \mid 0 \text{ exposures}) = \frac{1}{4}$$

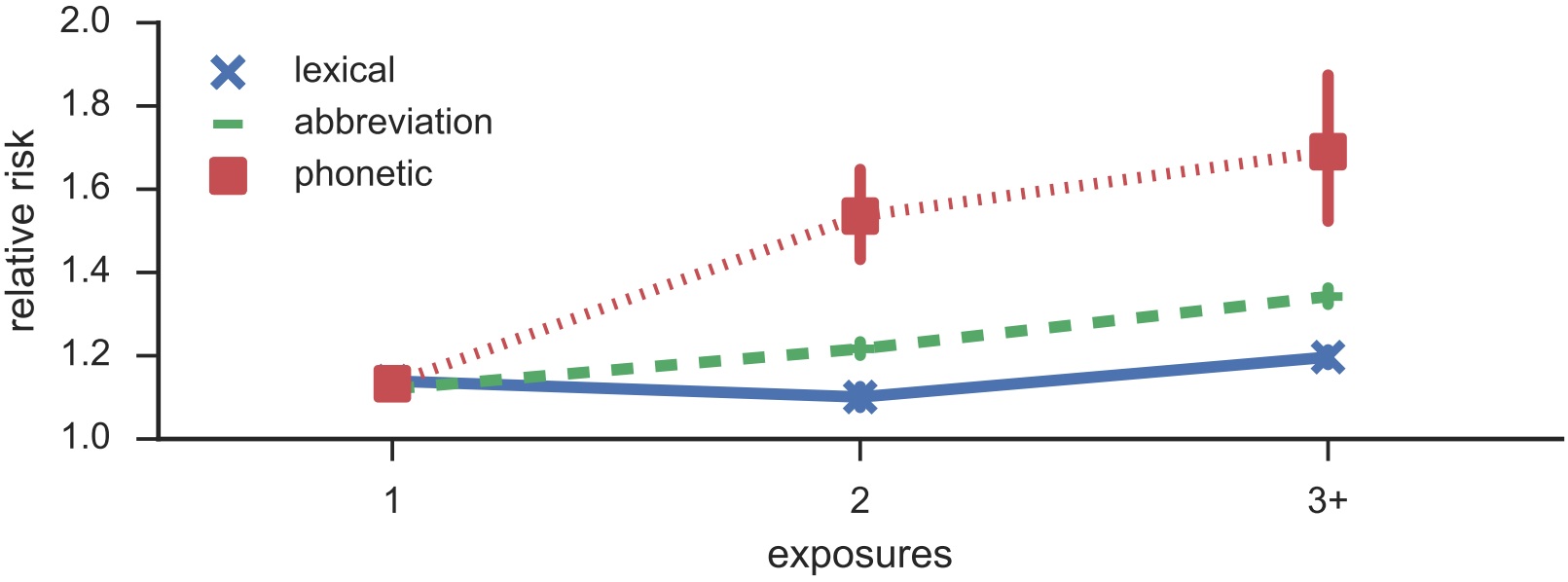
$$P(\text{infection} \mid \geq 1 \text{ exposures}) = \frac{2}{3}$$

- Randomized data

$$P(\text{infection} \mid 0 \text{ exposures}) = \frac{2}{4}, \frac{1}{4}, \ldots$$

$$P(\text{infection} \mid \geq 1 \text{ exposures}) = \frac{1}{2}, \frac{2}{3}, \ldots$$

Diffusion of innovations



- Relative risk: likelihood of infection given exposure, divided by likelihood under random shuffling
- Rel. risk > 1: evidence of non-random contagion.
- For phonetic variables, risk increases with multiple exposures, a characteristic of **complex contagion**.

Hypotheses

- **H1**: Language change is transmitted across social media networks.
- **H2**: Strong ties are better conduits of language change.
- **H3**: Geographically local social network ties are better conduits of language change.

Hypotheses

- **H1**: Language change is transmitted across social media networks.
- **H2**: Strong ties are better conduits of language change.
- **H3**: Geographically local social network ties are better conduits of language change.

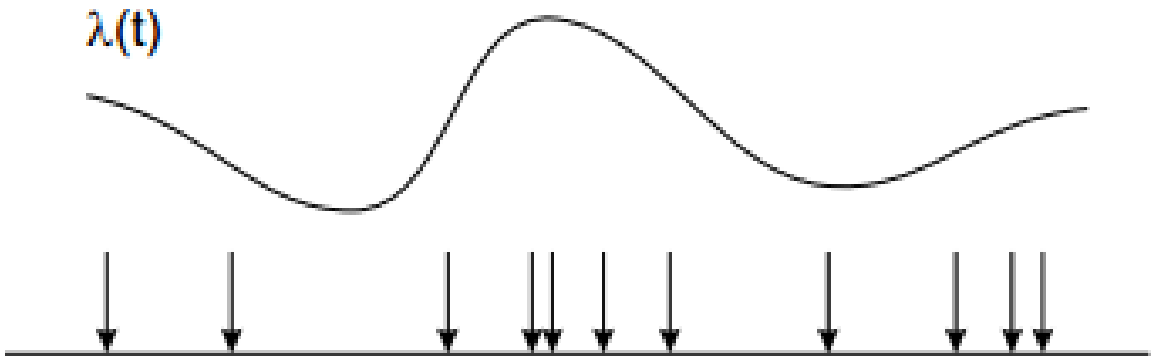
The Poisson process

- 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

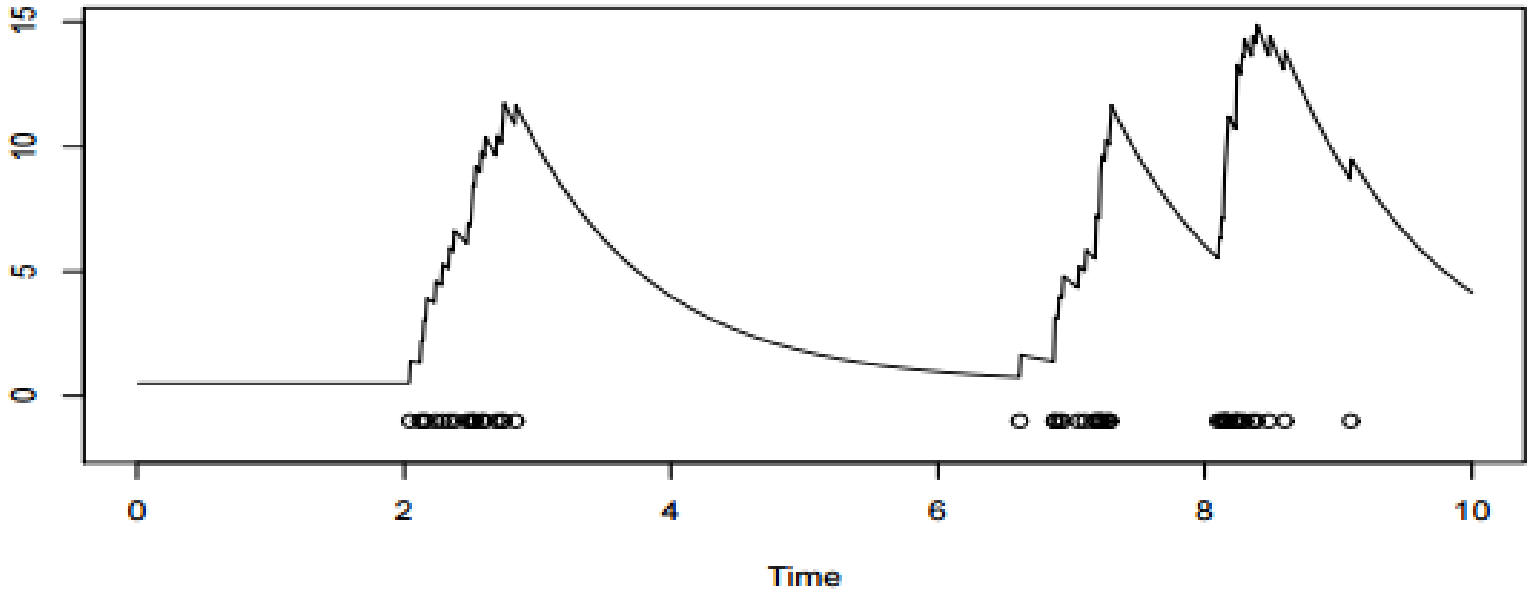
A Poisson process in which the intensity function depends on the history (Hawkes, 1971)

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

where,

- the time kernel κ decays exponentially with $t - t_n$;
- μ is the **base rate**;
- α captures the degree of self-excitation.

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,

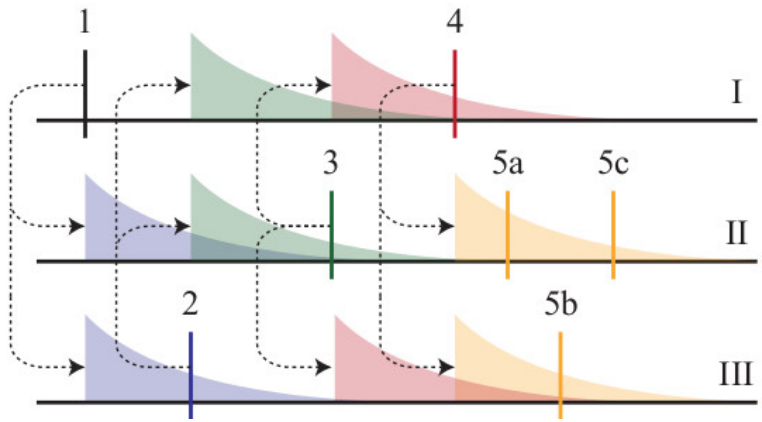
$$\lambda_m(t) = \mu_m + \sum_{t_n < t} \alpha_{m_n \to m} \kappa(t - t_n), \qquad (4)$$

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.



Parametric Hawkes process

The infection parameters are a linear function of shared features of each pair of individuals,

$$\alpha_{m_1 \to m_2} = \boldsymbol{\theta}^\top \boldsymbol{f}(m_1 \to m_2). \tag{5}$$

- We now need estimate only $\#|\theta|$ parameters, rather than M^2.
- We can use features to test hypotheses about what types of dyads are influential.

Features

F1:	**Self-excitation**	$\delta(m_1 = m_2)$
F2:	**Social network**	$(m_1, m_2) \in E$
F3:	**Locality**	$(m_1, m_2) \in E$ and $\text{loc}(m_1) = \text{loc}(m_2)$
F4:	**Tie strength**	$(m_1, m_2) \in E$ and (m_1, m_2) is a strong tie

Measuring tie strength

Mutual friends

$$mf(i,j) = \#|\{k : k \in \Gamma(i) \cap \Gamma(j)\}| \qquad (6)$$

Measuring tie strength

Mutual friends

$$mf(i,j) = \#|\{k : k \in \Gamma(i) \cap \Gamma(j)\}| \tag{6}$$

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

$$aa(i,j) = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log \#|\Gamma(k)|} \tag{7}$$

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

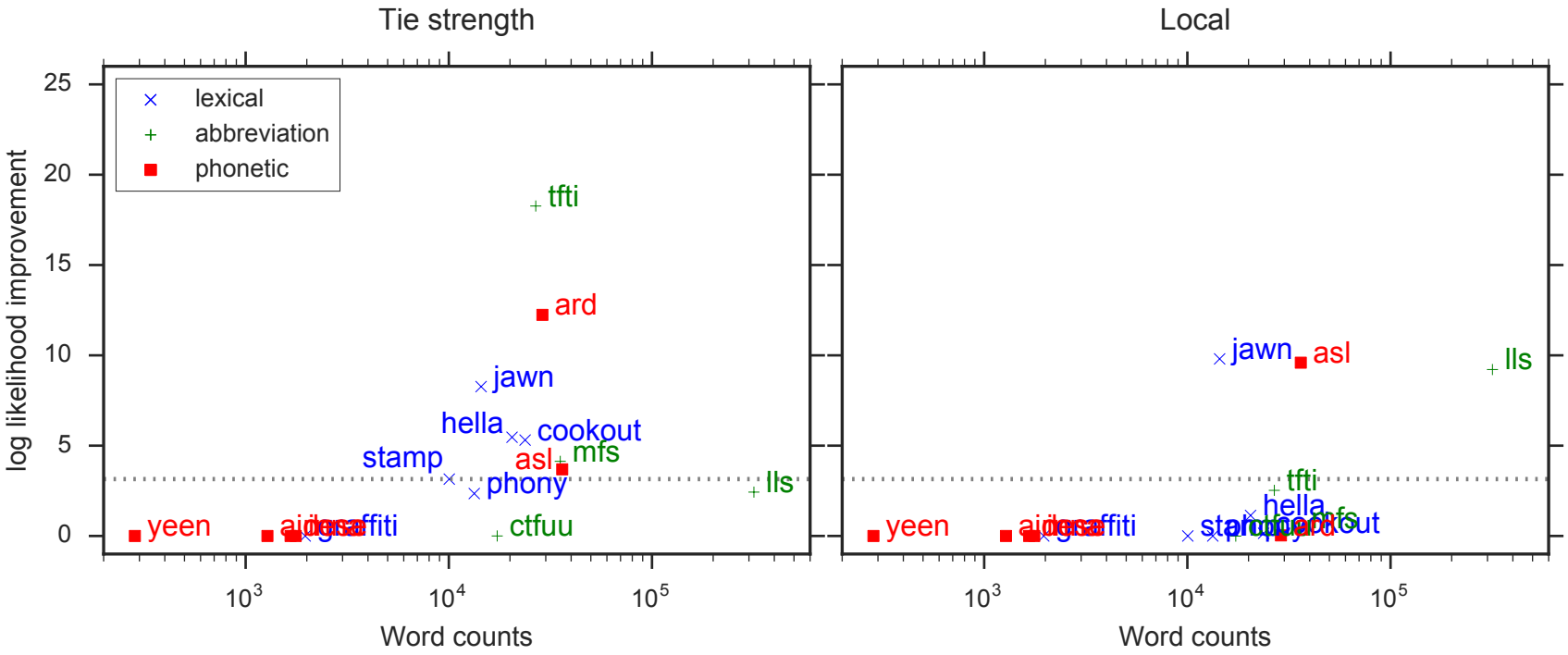
Hypothesis testing

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 strong ties better conduits of language change?

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

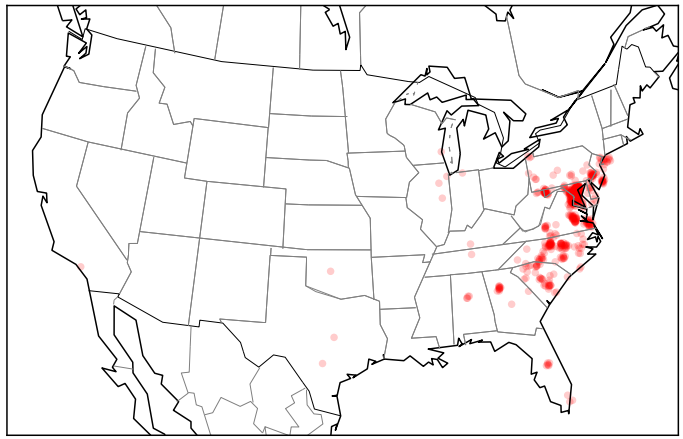
Goodness of fit



- Tie strength feature improves model fit for many words, evidence for H2.
- Geography locality rarely improves model fit, contradicting H3.

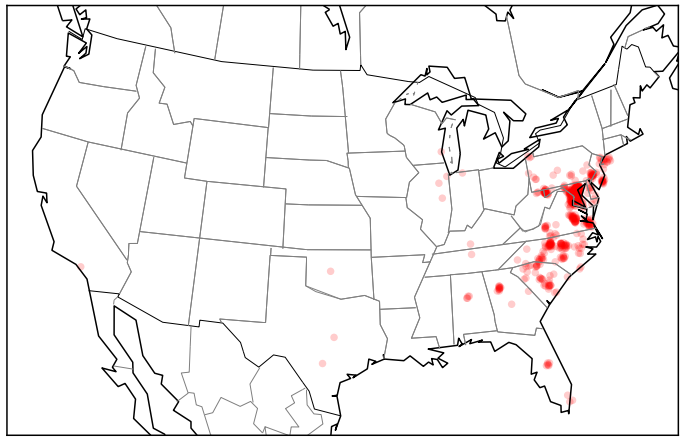
Why are social media variables so local?

No support for the hypothesis that exposures from local ties are especially influential.



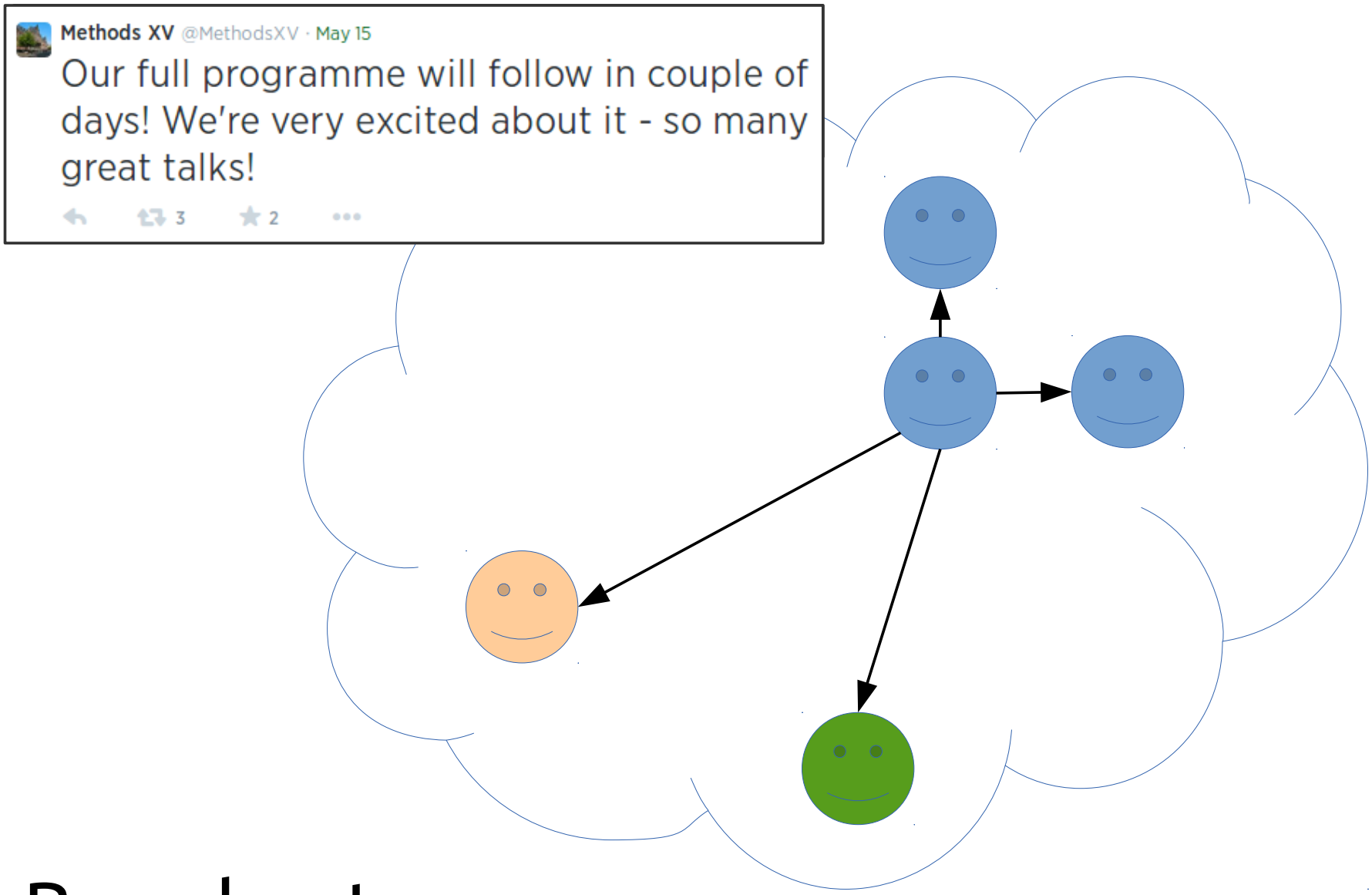
Why are social media variables so local?

No support for the hypothesis that exposures from local ties are especially influential.

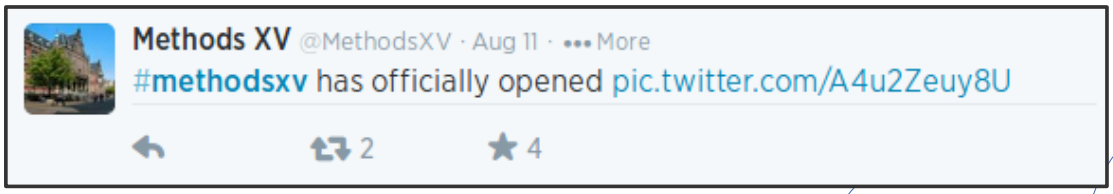


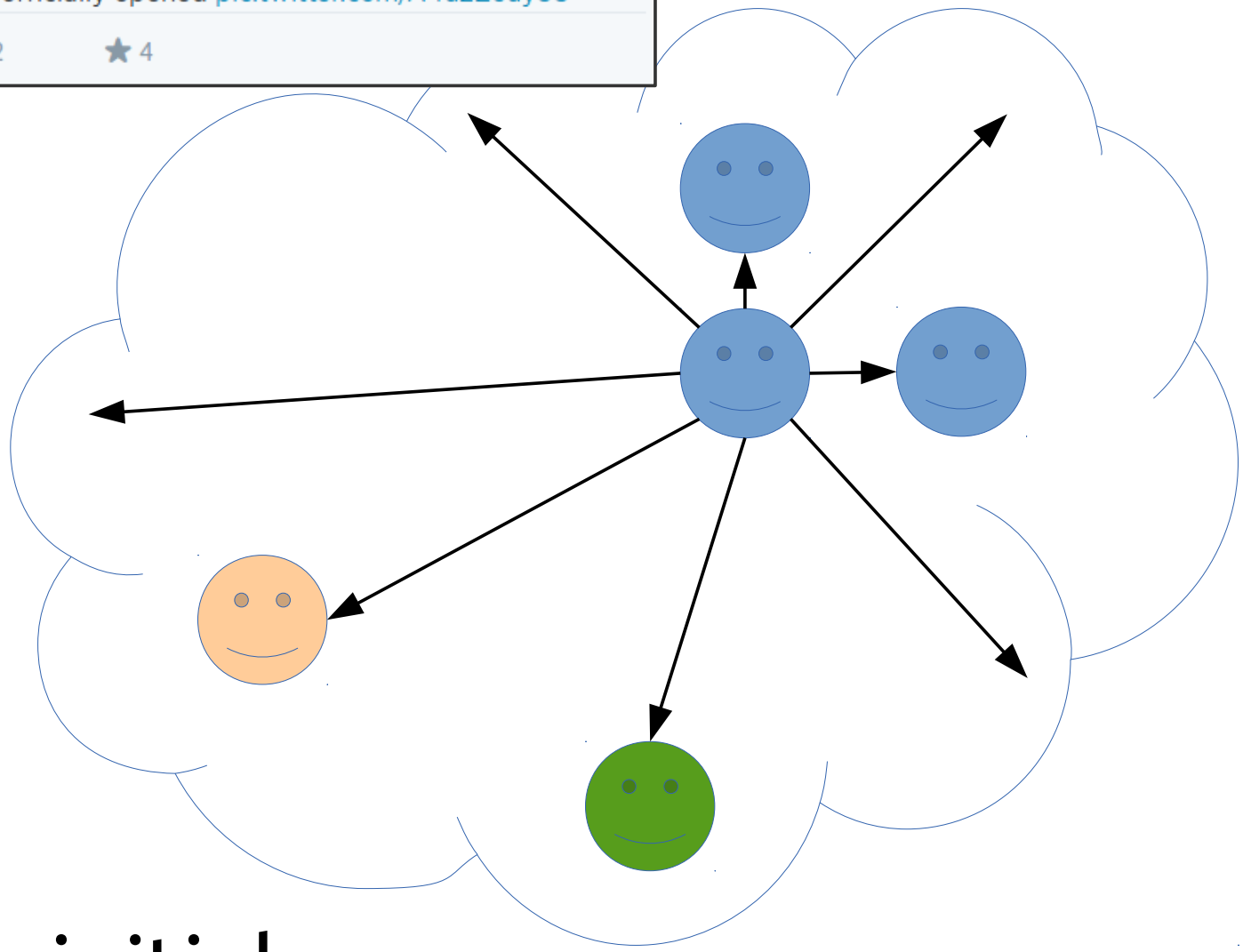
But! Most exposures are local. Two reasons:

1. Most ties are local
2. Audience design (Pavalanathan & Eisenstein, 2015)


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

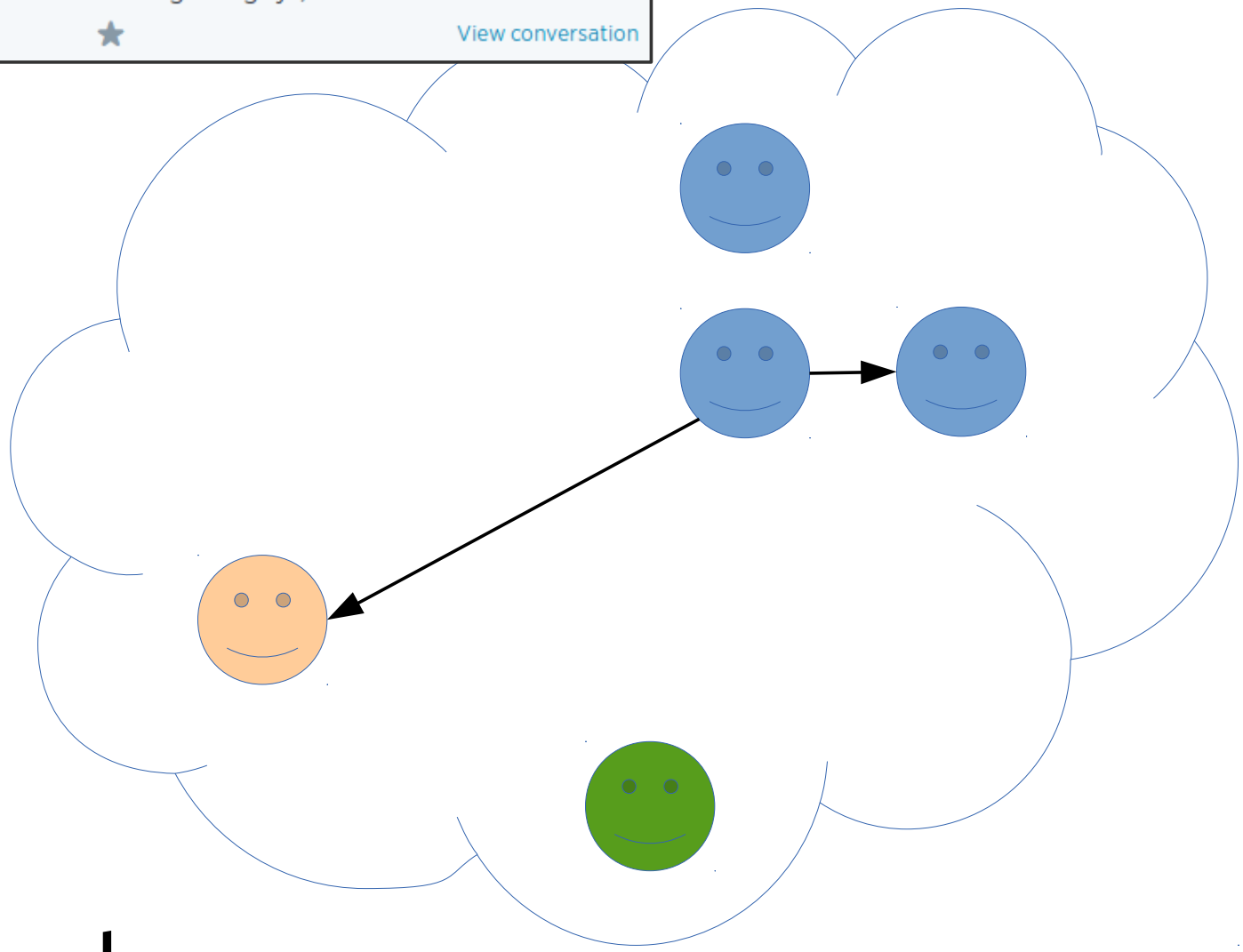
Broadcast





Hashtag-initial


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

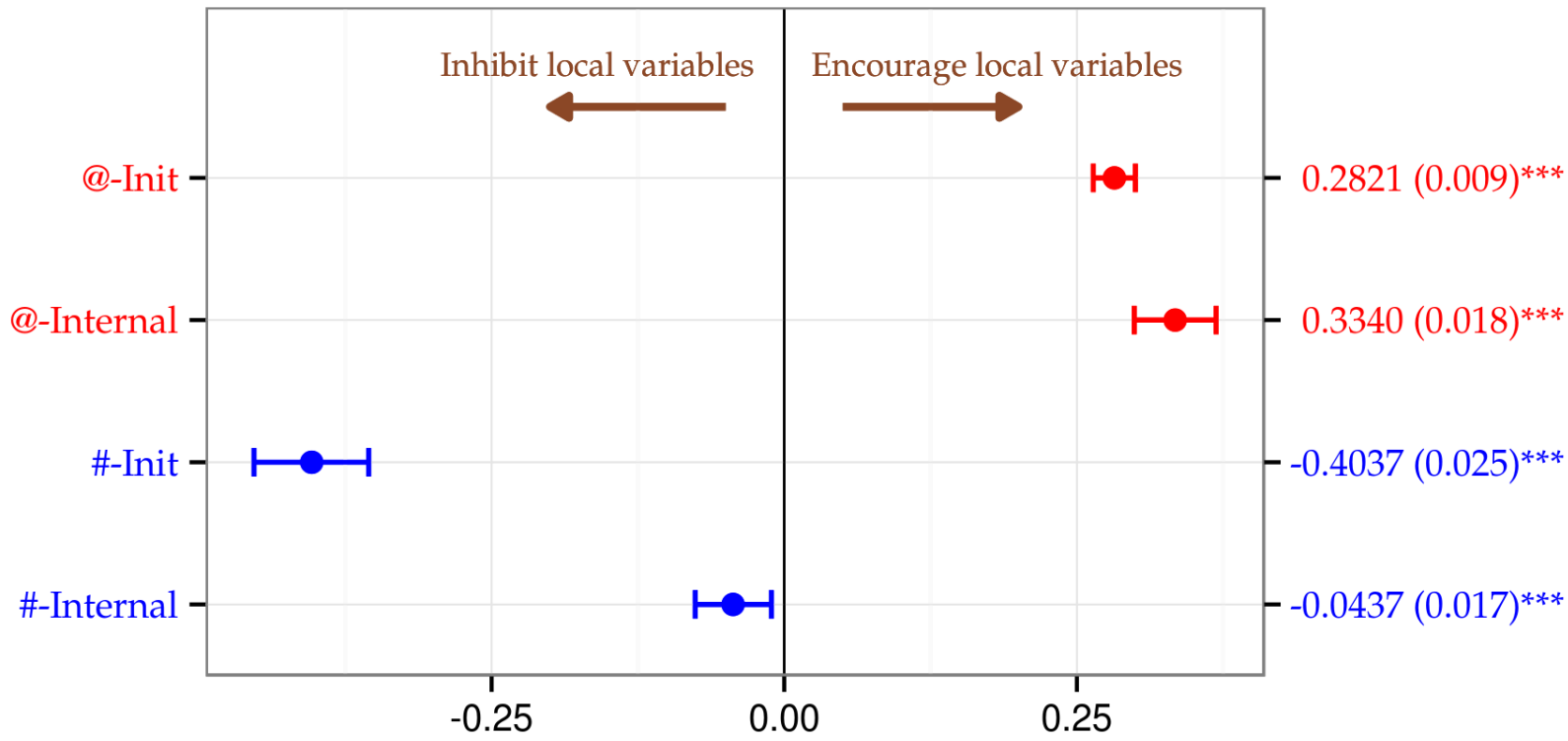


Addressed

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 $\rightarrow$ 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.

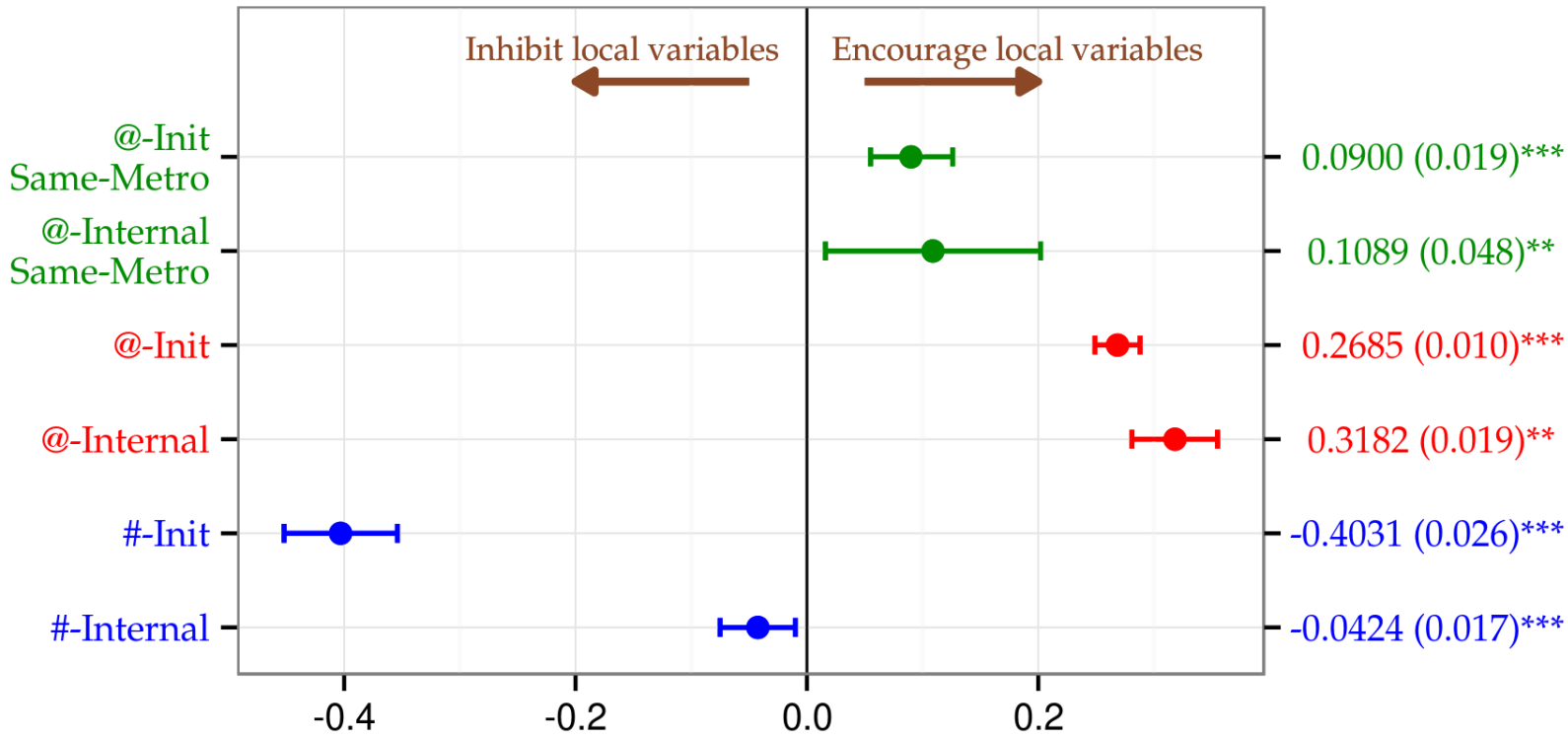
Distinguishing local ties

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.

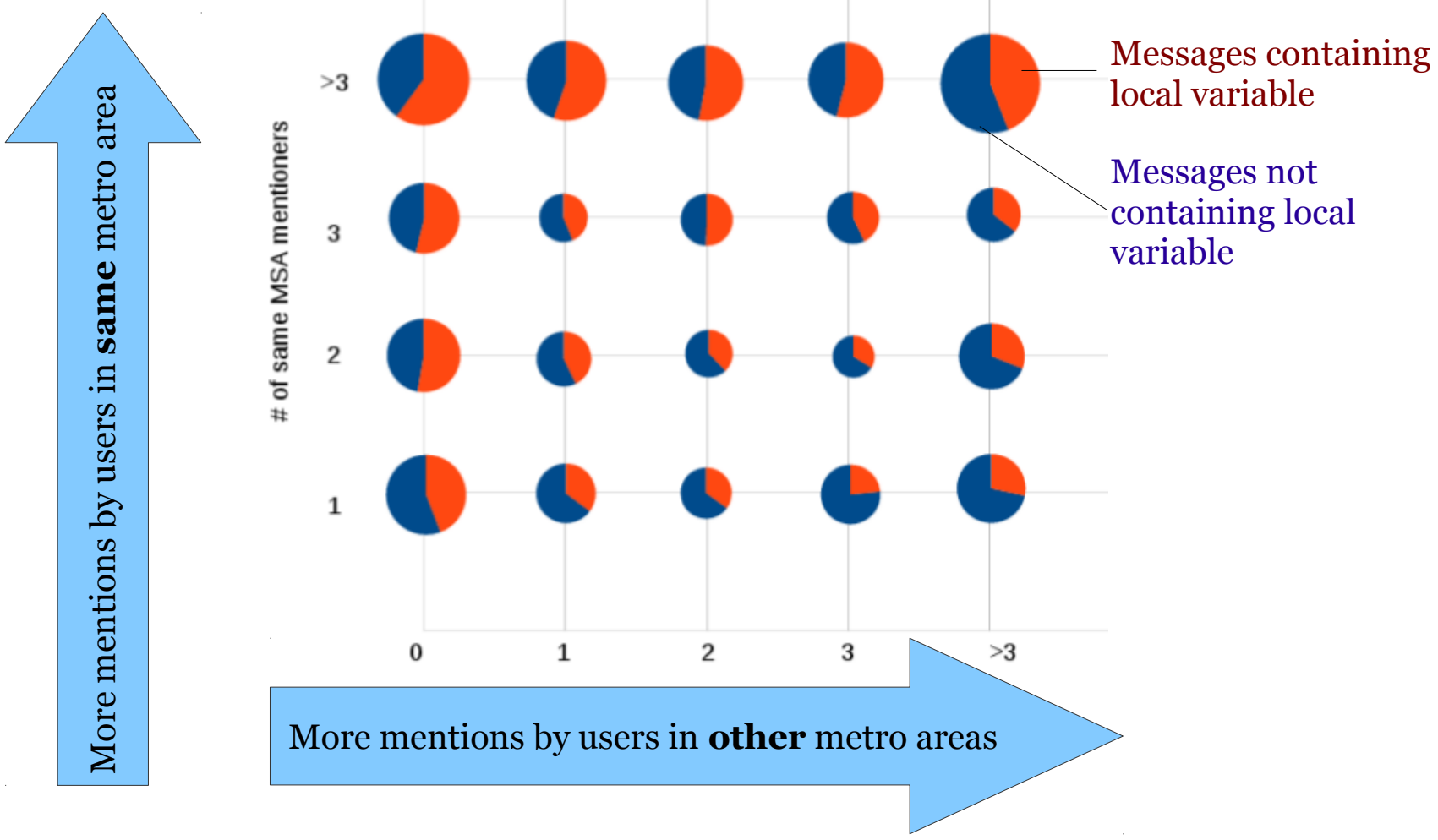
The social network lets us impute the locations of unknown users from the 1-2% of users who reveal their GPS! (Sadilek et al., 2012)

Local audience $\rightarrow$ less standard language



Local ties make non-standard language even more likely.

Local audience $\rightarrow$ less standard language



Next steps

- How to find geographically distinctive linguistic variables? (Nguyen & Eisenstein, 2017)
- What is the effect of incomplete data on estimates of linguistic influence?
- What is the impact of technological mediation on online language?

Next steps

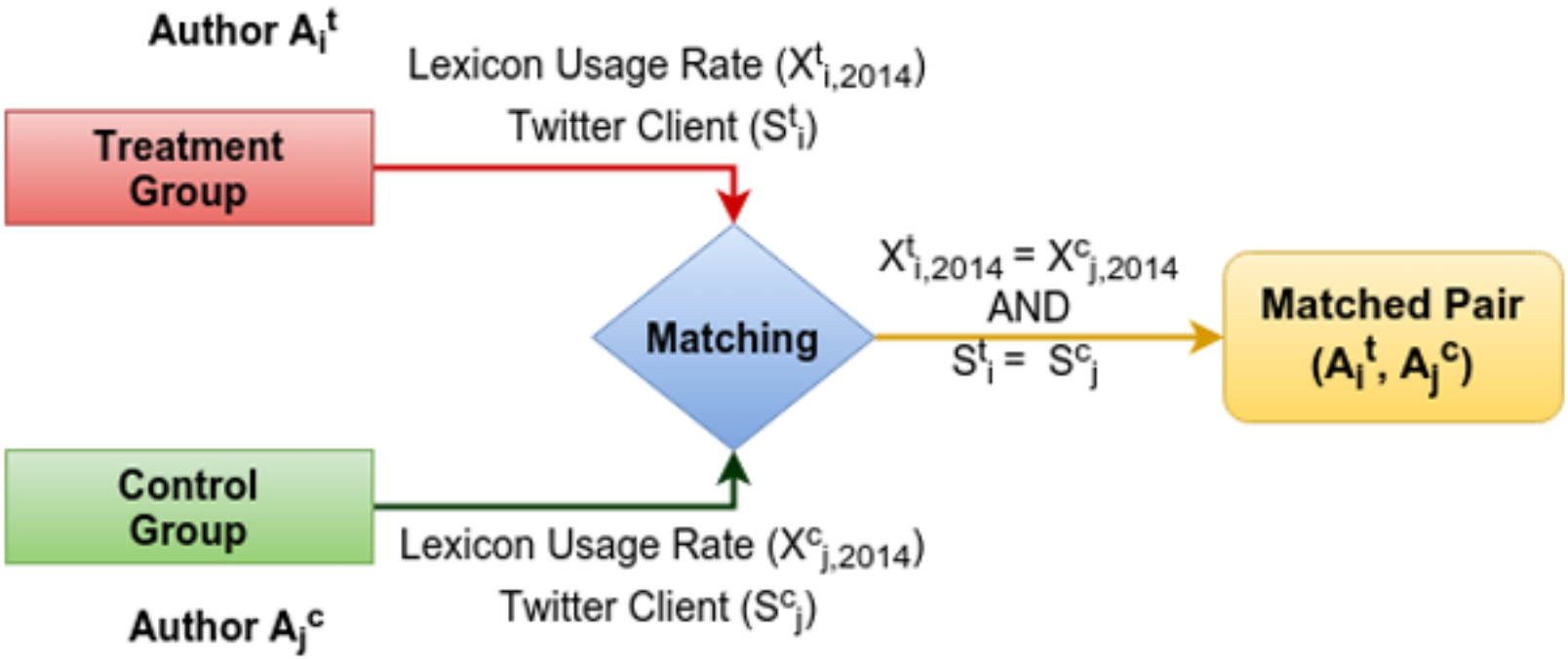
- How to find geographically distinctive linguistic variables? (Nguyen & Eisenstein, 2017)
- What is the effect of incomplete data on estimates of linguistic influence?
- **What is the impact of technological mediation on online language?**

Emojis vs emoticons

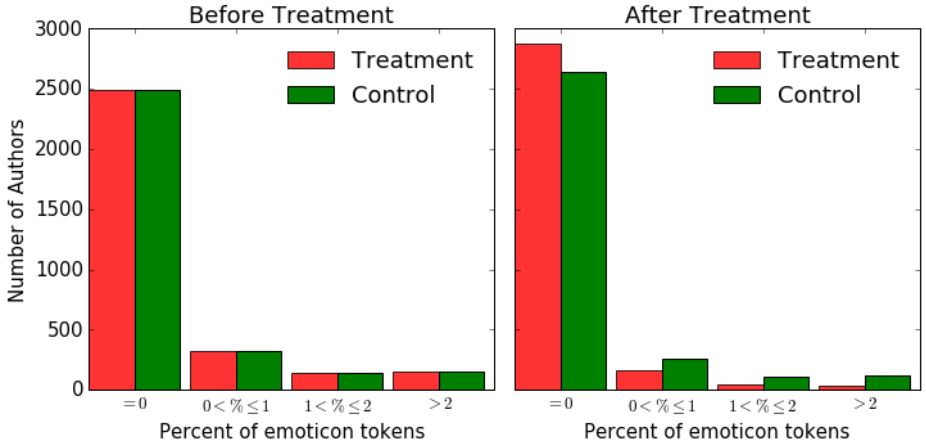
Pavalanathan & Eisenstein (2016): How has the introduction of emojis impacted the use of non-standard language?

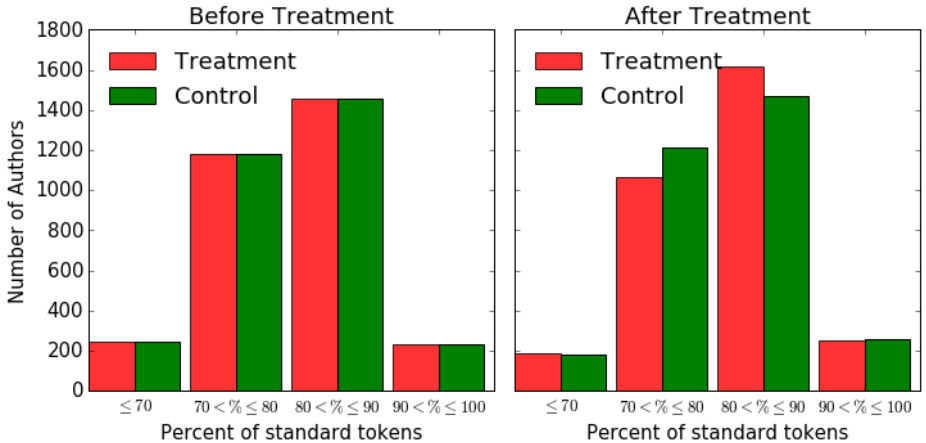
Causal inference design

- **Treatment**: adoption of emojis
- **Outcome**: frequency of emoticons and other non-standard tokens



Emojis vs emoticons





After treatment, emoji adopters use:

- significantly fewer emoticons ($0.40\% \rightarrow 0.12\%$)
- significantly more standard tokens ($85.2\% \rightarrow 86.2\%$)

Which emoticons?

Nose	:-) :-/
Horizontal	O_O -_-
Equal sign eyes	=) =D
Wink eyes	;-) ;D
Tongue	:P ;P
Slant mouth	:/ :-
Smiling mouth	:) ;)
Laughing mouth	:D ;-D
'O' mouth	:O :-O
Tears	:') :'(
Reversed	(-: (:

Which emoticons?

Nose	:-) :-/
Horizontal	O_O -_-
Equal sign eyes	=) =D
Wink eyes	;-) ;D
Tongue	:P ;P
Slant mouth	:/ :-
Smiling mouth	:) ;)
Laughing mouth	:D ;-D
'O' mouth	:O :-O
Tears	:') :'(
Reversed	(-: (:

Language variation: a challenge for NLP

"I would like to believe he's **sick** rather than just mean and evil."

Language variation: a challenge for NLP

"I would like to believe he's **sick** rather than just mean and evil."

"You could've been getting down to this **sick** beat."

(Yang & Eisenstein, 2017)

Personalization by ensemble

- Goal: personalized conditional likelihood, $P(y \mid \boldsymbol{x}, a)$, where a is the author.
- Problem: We have labeled examples for only a few authors.

Personalization by ensemble

- Goal: personalized conditional likelihood, $P(y \mid \boldsymbol{x}, a)$, where a is the author.
- Problem: We have labeled examples for only a few authors.
- **Personalization ensemble**

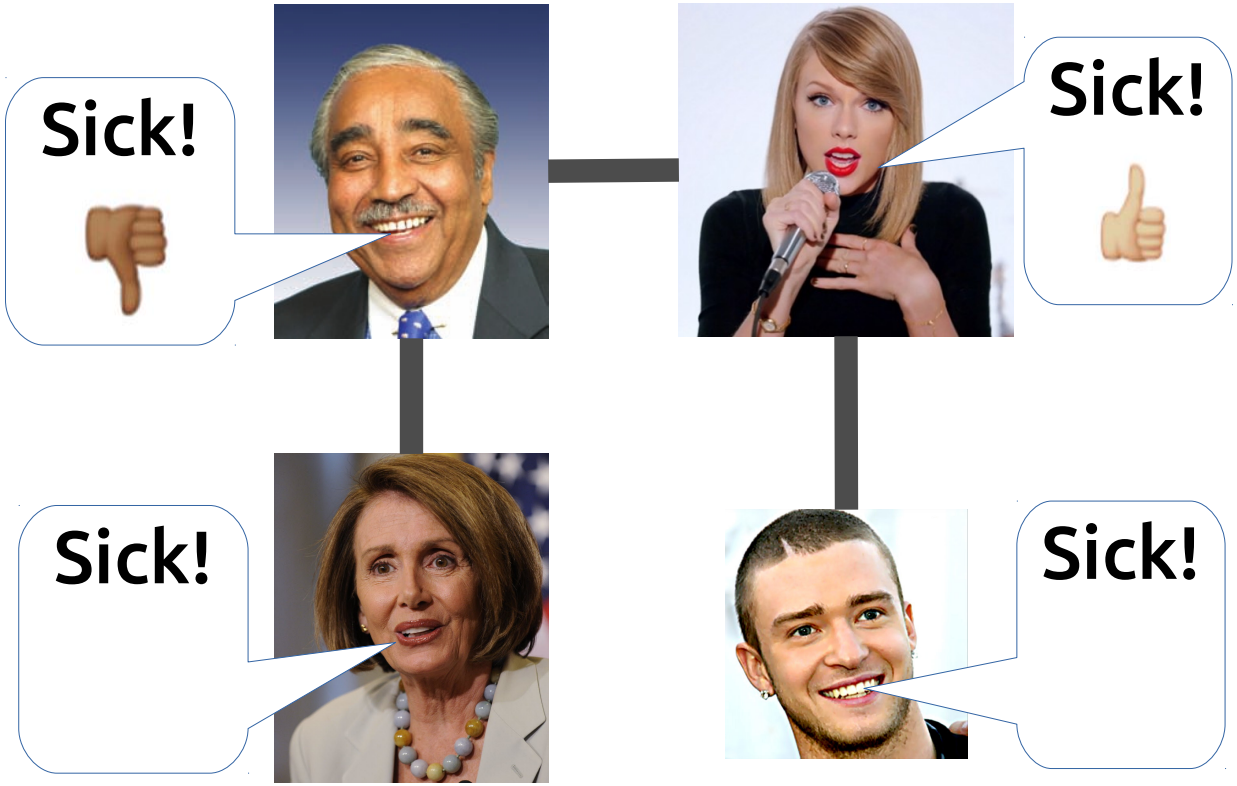
$$P(y \mid \boldsymbol{x}, a) = \sum_k P_k(y \mid \boldsymbol{x}) \pi_a(k)$$

  - $P_k(y \mid \boldsymbol{x})$ is a basis model
  - $\pi_a(\cdot)$ are the ensemble weights for author a

Homophily to the rescue?

Labeled data

Unlabeled data



Are language styles **assortative** on the social network?

Evidence for linguistic homophily

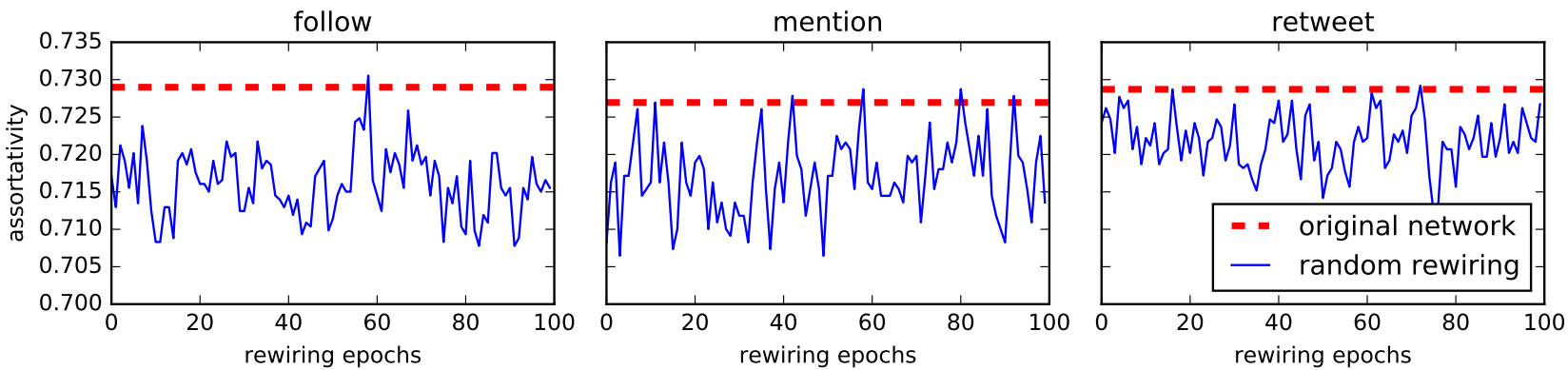
Pilot study: is classifier accuracy **assortative** on the Twitter social network?

$$\text{assort}(G) = \frac{1}{\#|G|} \sum_{(i,j) \in G} \delta(y_i = \hat{y}_i)\delta(y_j = \hat{y}_j) + \delta(y_i \neq \hat{y}_i)\delta(y_j \neq \hat{y}_j)$$

Evidence for linguistic homophily

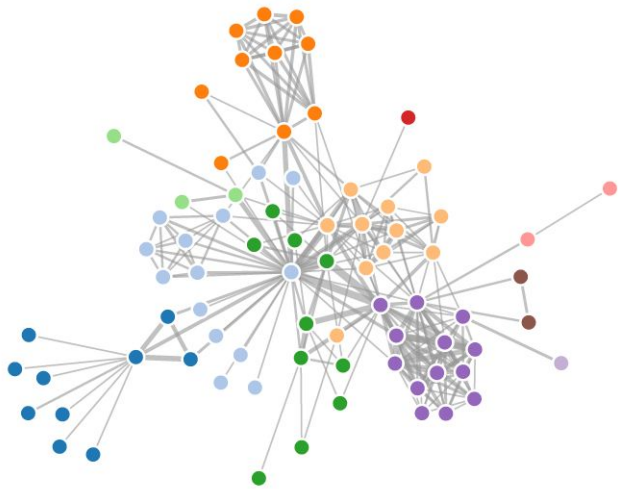
Pilot study: is classifier accuracy **assortative** on the Twitter social network?

$$\text{assort}(G) = \frac{1}{\#|G|} \sum_{(i,j)\in G} \delta(y_i = \hat{y}_i)\delta(y_j = \hat{y}_j) + \delta(y_i \neq \hat{y}_i)\delta(y_j \neq \hat{y}_j)$$



Network-driven personalization

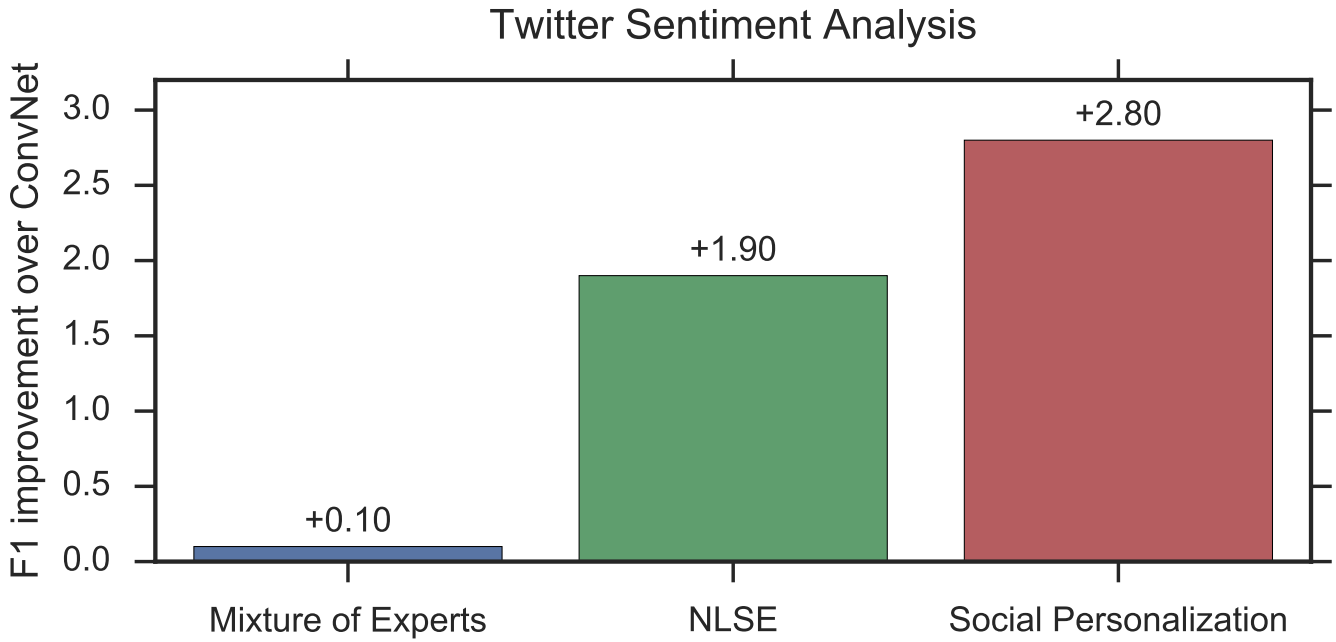
- For each author, estimate a **node embedding** $\boldsymbol{e}_a$ (Tang et al., 2015).
- Nodes who share neighbors get similar embeddings.



$$\boldsymbol{\pi}_a = \text{SoftMax}(f(\boldsymbol{e}_a))$$

$$P(y \mid \boldsymbol{x}, a) = \sum_{k=1}^{K} P_k(y \mid \boldsymbol{x}) \pi_a(k)$$

Results



Improvements over ConvNet baseline:

- +2.8% on Twitter Sentiment Analysis
- +2.7% on Ciao Product Reviews

NLSE is prior state-of-the-art (Astudillo et al., 2015).

Variable sentiment words

	More positive	More negative
1	banging loss fever broken fucking	dear like god yeah wow
2	chilling cold ill sick suck	satisfy trust wealth strong lmao
3	ass damn piss bitch shit	talent honestly voting win clever
4	insane bawling fever weird cry	lmao super lol haha hahaha
5	ruin silly bad boring dreadful	*lovatics* wish *beliebers arianators kendall*

Conclusions

- Internet media continues to create new social configurations, new subcultures, and new communicative affordances.
- New opportunities for studying the micro-foundations of language change.
- A future of blurred lines: online vs IRL, human control vs autonomy, text vs speech.

Acknowledgments

- **Collaborators**: Ming-Wei Chang, Fernando Diaz, Naman Goyal, Rahul Goel, Vinodh Krishnan, John Paparrizos, Umashanthi Pavalanathan, Sandeep Soni, Hanna Wallach, Yi Yang
- **Sponsors**: National Science Foundation, Air Force Office of Scientific Research, National Institutes for Health, Microsoft Research.

References I

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

Astudillo, R. F., Amir, S., Lin, W., Silva, M., & Trancoso, I. (2015). Learning word representations from scarce and noisy data with embedding sub-spaces. In *Proceedings of the Association for Computational Linguistics (ACL)*, Beijing.

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.

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

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).

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

Goel, R., Soni, S., Goyal, N., Paparrizos, J., Wallach, H., Diaz, F., & Eisenstein, J. (2016). The social dynamics of language change in online networks. In *The International Conference on Social Informatics (SocInfo)*.

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

Huang, Y., Guo, D., Kasakoff, A., & Grieve, J. (2016). Understanding us regional linguistic variation with twitter data analysis. *Computers, Environment and Urban Systems*, *59*, 244–255.

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

Nguyen, D. & Eisenstein, J. (2017). A kernel independence test for geographical language variation. *Computational Linguistics*, *in press*.

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

Pavalanathan, U. & Eisenstein, J. (2016). More emojis, less :) The competition for paralinguistic functions in microblog writing. *First Monday*, *22*(11).

References II

Sadilek, A., Kautz, H., & Bigham, J. P. (2012). Finding your friends and following them to where you are. In *Proceedings of the Conference on Web Search and Data Mining (WSDM)*, (pp. 723–732).

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., & Mei, Q. (2015). Line: Large-scale information network embedding. In *Proceedings of the Conference on World-Wide Web (WWW)*, (pp. 1067–1077).

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

Yang, Y. & Eisenstein, J. (2017). Overcoming language variation in sentiment analysis with social attention. *Transactions of the Association for Computational Linguistics (TACL)*.