Machine learning for computational social science

Jacob Eisenstein

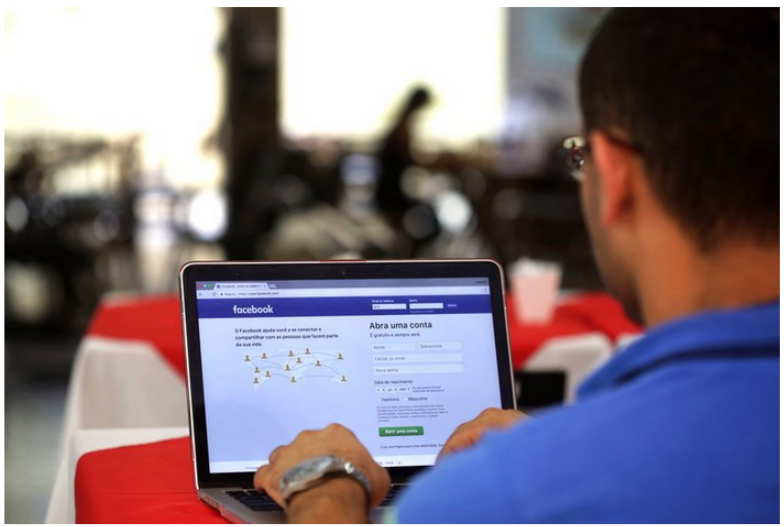
Georgia Institute of Technology

February 15, 2018

TECHNOLOGY

Facebook to Let Users Rank Credibility of News

By SHEERA FRENKEL and SAPNA MAHESHWARI JAN. 19, 2018



Twitter admits far more Russian bots posted on election than it had disclosed

Company says it removed more than 50,000 accounts and reported them to investigators, marking latest upward revision of figures

Twitter has admitted that more than 50,000 Russia-linked accounts used its service to post automated material about the 2016 US election - a far greater number than previously disclosed.

The social and political domains are now computationally mediated.

- **New challenges**: echo chambers, bots, viral hoaxes, hate speech
- **New opportunities**: to measure and understand social phenomena; to address social challenges at scale

These changes motivate the emerging discipline of **computational social science**.

Routes to computational social science

Ready-made:
high-quality software that can be used for a range of social science problems
(R, AllenNLP, gephi)

Routes to computational social science

Ready-made:
high-quality software that can be used for a range of social science problems (R, AllenNLP, gephi)

Custom builds: bespoke computational solutions for specific social science research problems

Three short pieces

- Exploring the construction of social meaning in networks

  (Krishnan & Eisenstein, 2015; Pavalanathan et al., 2017)

- Operationalizing sociocultural influence from spatiotemporal cascades

  (Eisenstein et al., 2014; Goel et al., 2016; Soni & Eisenstein, 2018)

- Measuring the causal impact of efforts to combat hate speech

  (Chandrasekharan et al., 2018)

Three short pieces

- **Exploring the construction of social meaning in networks**

  (Krishnan & Eisenstein, 2015; Pavalanathan et al., 2017)

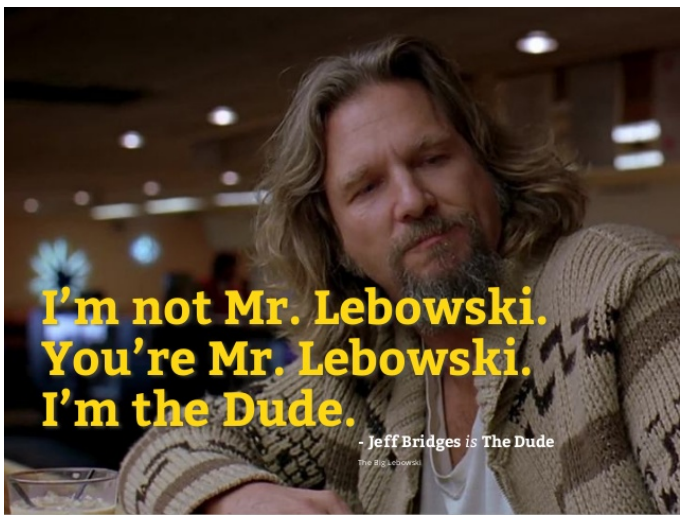
- Operationalizing sociocultural influence from spatiotemporal cascades

  (Eisenstein et al., 2014; Goel et al., 2016; Soni & Eisenstein, 2018)

- Measuring the causal impact of efforts to combat hate speech

  (Chandrasekharan et al., 2018)

Social meaning in social networks

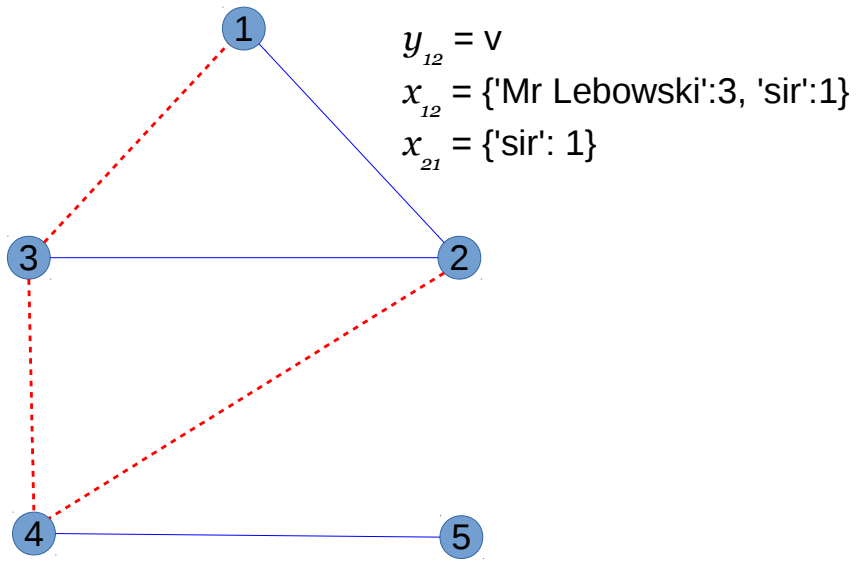


- How do **address terms** like Mr. Lebowski and dude create social meaning?
- How are social relationships arranged on a network?
- Can we leverage text and networks to make more accurate inferences about interpersonal relationships?

Formulation as a signed social network

Variables:

- network structure $G = \{(i, j)\},\ i < j$
- linguistic features $\boldsymbol{x}_{i \rightarrow j} \in \mathbb{N}^V$
- **latent** edge labels $y_{ij} \in \mathcal{Y}$

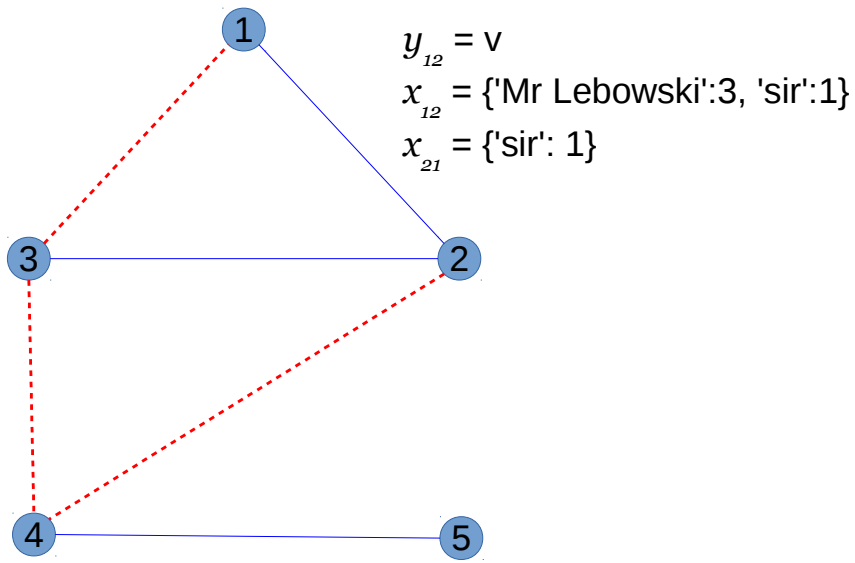


Formulation as a signed social network

Modeling assumption: each edge label indexes a distribution over address terms

$$\boldsymbol{x}_{i \rightarrow j} \mid y_{ij} \sim \text{Multinomial}(\boldsymbol{\theta}_{y_{ij}})$$
$$\boldsymbol{x}_{i \leftarrow j} \mid y_{ij} \sim \text{Multinomial}(\boldsymbol{\theta}_{y_{ij}}).$$

Estimating $\boldsymbol{\theta}$ gives the distribution over address terms for each edge type.

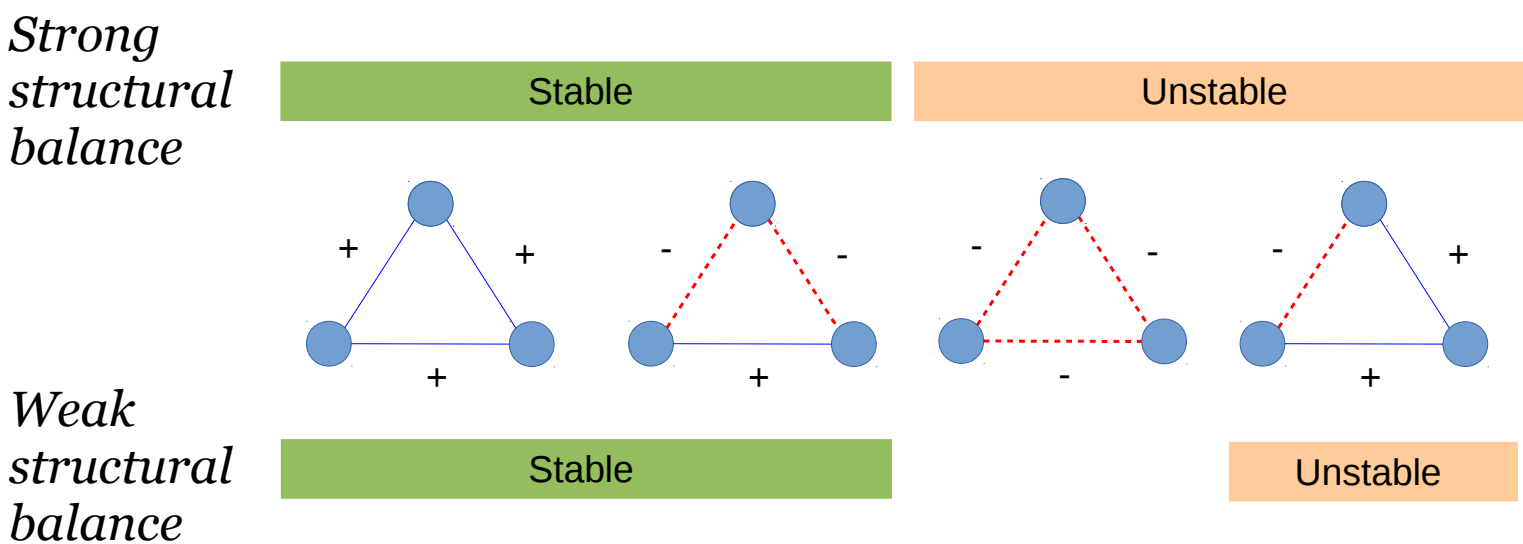


Stable and unstable label configurations

- So far, this is just a mixture model over dyads.
- Are some label configurations better than others?

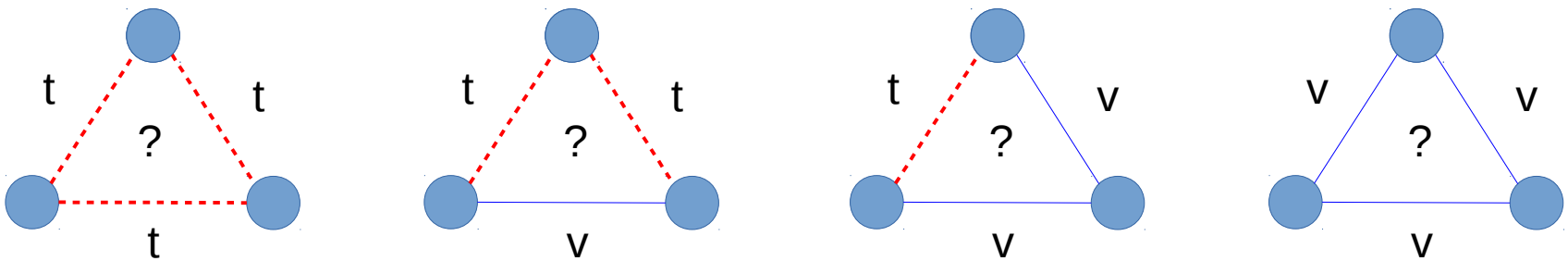
Stable and unstable label configurations

- So far, this is just a mixture model over dyads.
- Are some label configurations better than others?
- **Structural balance theory** describes networks of friend/enemy links, where signed triads may be stable or unstable:



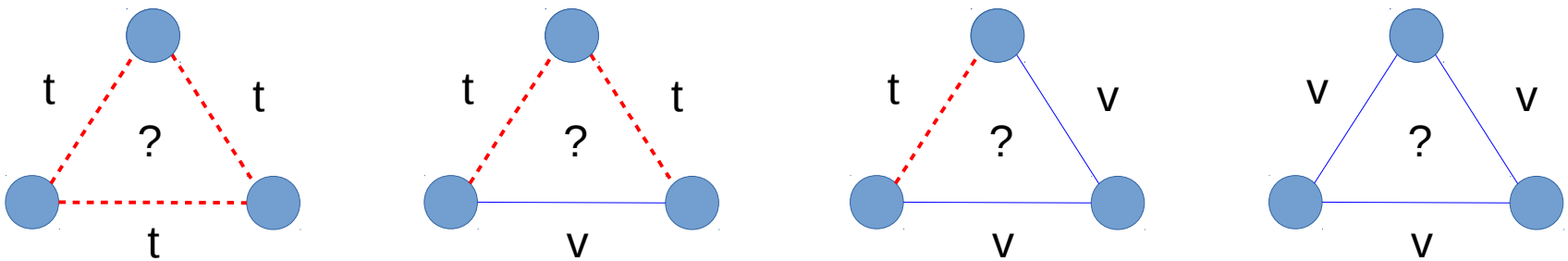
Network models with unknown parameters

What if the magnitude, and even the direction of the effect of each triad type is unknown?



Network models with unknown parameters

What if the magnitude, and even the direction of the effect of each triad type is unknown?



Structure induction: estimate weights on each triad type, using only unlabeled data.

Prior distribution over signed networks

Assume the prior factors over dyads and triads.

$$P(\boldsymbol{y}; G, \boldsymbol{\eta}, \boldsymbol{\beta}) = \frac{1}{Z(\boldsymbol{\eta}, \boldsymbol{\beta}; G)} \times \exp \sum_{\langle i,j \rangle \in G} \boldsymbol{\eta} \cdot \boldsymbol{f}(y_{ij}, i, j, G) \times \exp \sum_{\langle i,j,k \rangle \in \mathcal{T}(G)} \beta_{y_{ij}, y_{jk}, y_{ik}},$$

where,

- $\boldsymbol{f}(y_{ij}, i, j, G)$ is a set of dyad features, with associated weights $\boldsymbol{\eta}$;
- $\mathcal{T}(G)$ is the set of triads in the graph G;
- $\beta_{y_{ij}, y_{jk}, y_{ik}}$ scores the stability of a triad type.
- $Z(\boldsymbol{\eta}, \boldsymbol{\beta}; G)$ is a normalizing constant;

Prior distribution over signed networks

Assume the prior factors over dyads and triads.

$$P(\boldsymbol{y}; G, \boldsymbol{\eta}, \boldsymbol{\beta}) = \frac{1}{Z(\boldsymbol{\eta}, \boldsymbol{\beta}; G)} \times \exp \sum_{\langle i,j \rangle \in G} \boldsymbol{\eta} \cdot \boldsymbol{f}(y_{ij}, i, j, G) \times \exp \sum_{\langle i,j,k \rangle \in \mathcal{T}(G)} \beta_{y_{ij}, y_{jk}, y_{ik}},$$

where,

- $\boldsymbol{f}(y_{ij}, i, j, G)$ is a set of dyad features, with associated weights $\boldsymbol{\eta}$;
- $\mathcal{T}(G)$ is the set of triads in the graph G;
- $\beta_{y_{ij}, y_{jk}, y_{ik}}$ scores the stability of a triad type.
- $Z(\boldsymbol{\eta}, \boldsymbol{\beta}; G)$ is a normalizing constant;

Prior distribution over signed networks

Assume the prior factors over dyads and triads.

$$P(\boldsymbol{y}; G, \boldsymbol{\eta}, \boldsymbol{\beta}) = \frac{1}{Z(\boldsymbol{\eta}, \boldsymbol{\beta}; G)} \times \exp \sum_{\langle i,j \rangle \in G} \boldsymbol{\eta} \cdot \boldsymbol{f}(y_{ij}, i, j, G) \times \exp \sum_{\langle i,j,k \rangle \in \mathcal{T}(G)} \beta_{y_{ij}, y_{jk}, y_{ik}},$$

where,

- $\boldsymbol{f}(y_{ij}, i, j, G)$ is a set of dyad features, with associated weights $\boldsymbol{\eta}$;
- $\mathcal{T}(G)$ is the set of triads in the graph G;
- $\beta_{y_{ij}, y_{jk}, y_{ik}}$ scores the stability of a triad type.
- $Z(\boldsymbol{\eta}, \boldsymbol{\beta}; G)$ is a normalizing constant;

Prior distribution over signed networks

Assume the prior factors over dyads and triads.

$$P(\boldsymbol{y}; G, \boldsymbol{\eta}, \boldsymbol{\beta}) = \frac{1}{Z(\boldsymbol{\eta}, \boldsymbol{\beta}; G)} \times \exp \sum_{\langle i,j \rangle \in G} \boldsymbol{\eta} \cdot \boldsymbol{f}(y_{ij}, i, j, G) \times \exp \sum_{\langle i,j,k \rangle \in \mathcal{T}(G)} \beta_{y_{ij}, y_{jk}, y_{ik}},$$

where,

- $\boldsymbol{f}(y_{ij}, i, j, G)$ is a set of dyad features, with associated weights $\boldsymbol{\eta}$;
- $\mathcal{T}(G)$ is the set of triads in the graph G;
- $\beta_{y_{ij}, y_{jk}, y_{ik}}$ scores the stability of a triad type.
- $Z(\boldsymbol{\eta}, \boldsymbol{\beta}; G)$ is a normalizing constant;

Prior distribution over signed networks

Assume the prior factors over dyads and triads.

$$P(\boldsymbol{y}; G, \boldsymbol{\eta}, \boldsymbol{\beta}) = \frac{1}{Z(\boldsymbol{\eta}, \boldsymbol{\beta}; G)} \times \exp \sum_{\langle i,j \rangle \in G} \boldsymbol{\eta} \cdot \boldsymbol{f}(y_{ij}, i, j, G) \times \exp \sum_{\langle i,j,k \rangle \in \mathcal{T}(G)} \beta_{y_{ij}, y_{jk}, y_{ik}},$$

where,

- $\boldsymbol{f}(y_{ij}, i, j, G)$ is a set of dyad features, with associated weights $\boldsymbol{\eta}$;
- $\mathcal{T}(G)$ is the set of triads in the graph G;
- $\beta_{y_{ij}, y_{jk}, y_{ik}}$ scores the stability of a triad type.
- $Z(\boldsymbol{\eta}, \boldsymbol{\beta}; G)$ is a normalizing constant;

Complete model specification

$$P(\boldsymbol{y}, \boldsymbol{x} \mid G; \Theta, \boldsymbol{\beta}, \boldsymbol{\eta}) = P(\boldsymbol{x} \mid \boldsymbol{y}; \Theta) P(\boldsymbol{y} \mid G; \boldsymbol{\beta}, \boldsymbol{\eta})$$

- The prior factors across dyads and triads;
- The likelihood factors across dyads.

Complete model specification

$$P(\boldsymbol{y}, \boldsymbol{x} \mid G; \Theta, \boldsymbol{\beta}, \boldsymbol{\eta}) = P(\boldsymbol{x} \mid \boldsymbol{y}; \Theta) P(\boldsymbol{y} \mid G; \boldsymbol{\beta}, \boldsymbol{\eta})$$

- The **prior** factors across dyads and triads;
- The likelihood factors across dyads.

Complete model specification

$$P(\boldsymbol{y}, \boldsymbol{x} \mid G; \Theta, \boldsymbol{\beta}, \boldsymbol{\eta}) = P(\boldsymbol{x} \mid \boldsymbol{y}; \Theta) P(\boldsymbol{y} \mid G; \boldsymbol{\beta}, \boldsymbol{\eta})$$

- The prior factors across dyads and triads;
- The **likelihood** factors across dyads.

Complete model specification

$$P(\boldsymbol{y}, \boldsymbol{x} \mid G; \Theta, \boldsymbol{\beta}, \boldsymbol{\eta}) = P(\boldsymbol{x} \mid \boldsymbol{y}; \Theta) P(\boldsymbol{y} \mid G; \boldsymbol{\beta}, \boldsymbol{\eta})$$

Inference in this model answers several questions:

1. What is the relationship of each dyad?
2. How are social relationships expressed in language?
3. Are there structural regularities across networks?

Complete model specification

$$P(\boldsymbol{y}, \boldsymbol{x} \mid G; \Theta, \boldsymbol{\beta}, \boldsymbol{\eta}) = P(\boldsymbol{x} \mid \boldsymbol{y}; \Theta) P(\boldsymbol{y} \mid G; \boldsymbol{\beta}, \boldsymbol{\eta})$$

Inference in this model answers several questions:

1. What is the relationship of each dyad?
2. How are social relationships expressed in language?
3. Are there structural regularities across networks?

Complete model specification

$$P(\boldsymbol{y}, \boldsymbol{x} \mid G; \Theta, \boldsymbol{\beta}, \boldsymbol{\eta}) = P(\boldsymbol{x} \mid \boldsymbol{y}; \Theta) P(\boldsymbol{y} \mid G; \boldsymbol{\beta}, \boldsymbol{\eta})$$

Inference in this model answers several questions:

1. What is the relationship of each dyad?
2. How are social relationships expressed in language?
3. Are there structural regularities across networks?

Complete model specification

$$P(\boldsymbol{y}, \boldsymbol{x} \mid G; \Theta, \boldsymbol{\beta}, \boldsymbol{\eta}) = P(\boldsymbol{x} \mid \boldsymbol{y}; \Theta) P(\boldsymbol{y} \mid G; \boldsymbol{\beta}, \boldsymbol{\eta})$$

Inference in this model answers several questions:

1. What is the relationship of each dyad?
2. How are social relationships expressed in language?
3. Are there structural regularities across networks?

Application: a dataset of 617 movie scripts (Danescu-Niculescu-Mizil & Lee, 2011).

Intractability

- **Inference**: optimizing an objective over dyads and triads is NP-hard (West et al., 2014).

  We iteratively update a mean field relaxation, $Q(\boldsymbol{y}) = \prod_{\langle i,j \rangle \in G} q_{ij}(y_{ij})$.

Intractability

- **Inference**: optimizing an objective over dyads and triads is NP-hard (West et al., 2014).

  We iteratively update a mean field relaxation, $Q(\boldsymbol{y}) = \prod_{\langle i,j \rangle \in G} q_{ij}(y_{ij})$.

- **Learning**: computing the normalizing constant Z requires summing across exponentially many possible labelings.

  We apply noise-contrastive estimation (Gutmann & Hyvärinen, 2012) in the M-step to approximate gradients on the parameters $\boldsymbol{\theta}$, $\boldsymbol{\eta}$, and $\boldsymbol{\beta}$.

Validation

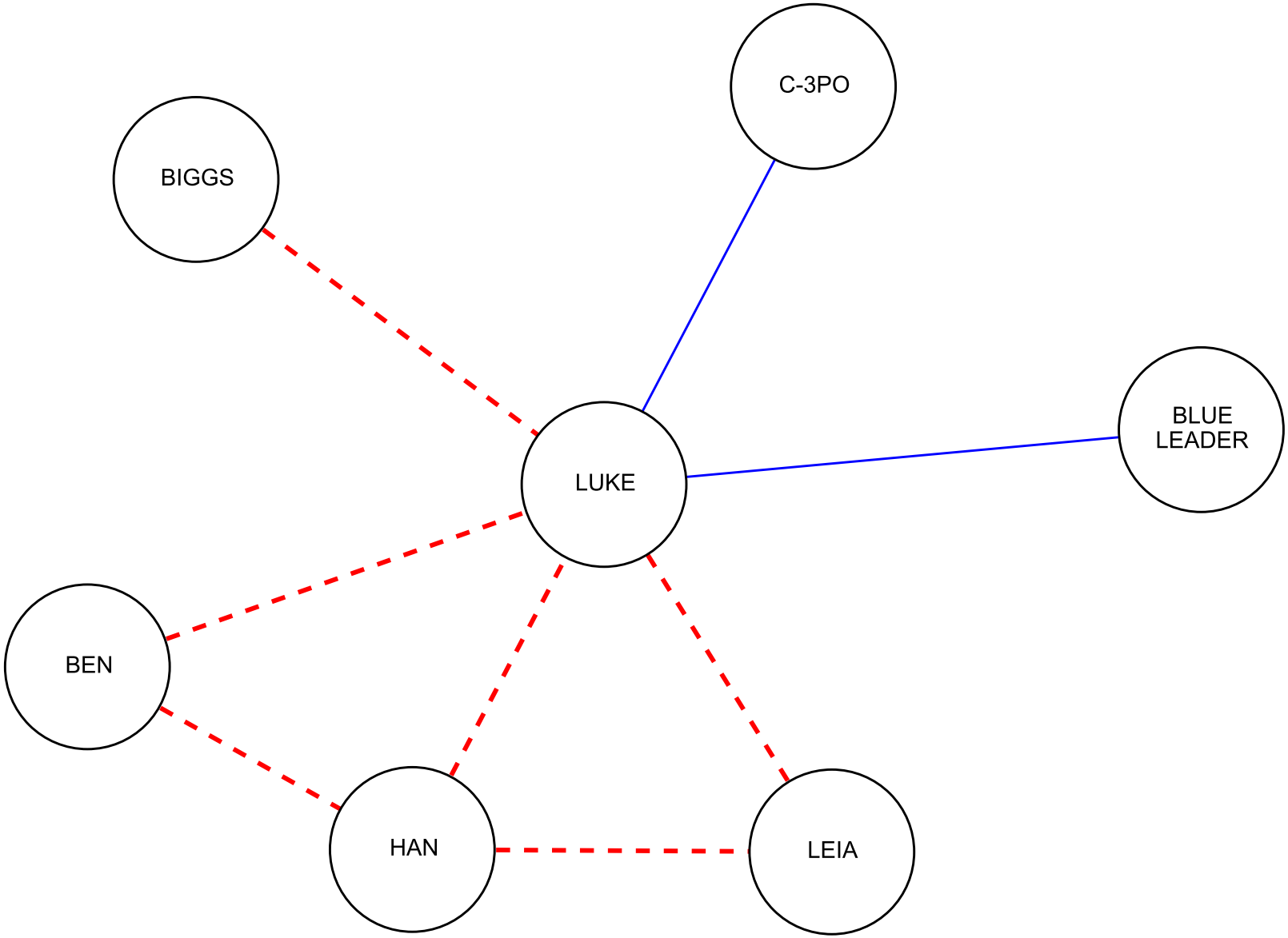
sir	FIRSTNAME
mr+LASTNAME	man
mr+FIRSTNAME	baby
mr	honey
miss+LASTNAME	darling
son	sweetheart
mister+FIRSTNAME	buddy
mrs	sweetie

Validation

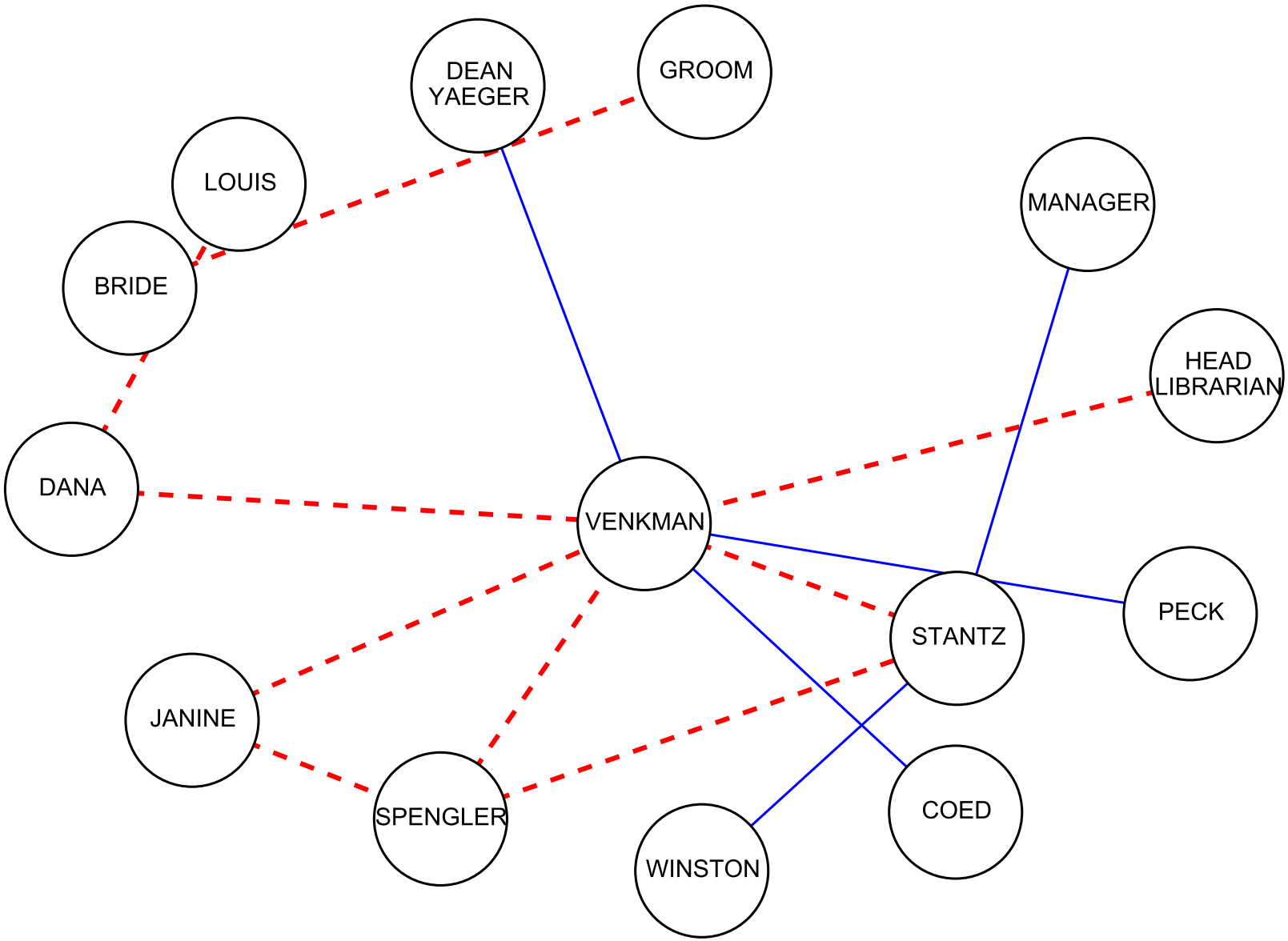
sir	FIRSTNAME
mr+LASTNAME	man
mr+FIRSTNAME	baby
mr	honey
miss+LASTNAME	darling
son	sweetheart
mister+FIRSTNAME	buddy
mrs	sweetie

- Raters found the intruder term in 73% of cases for the full model.
- ... versus 52% in the text-only model.

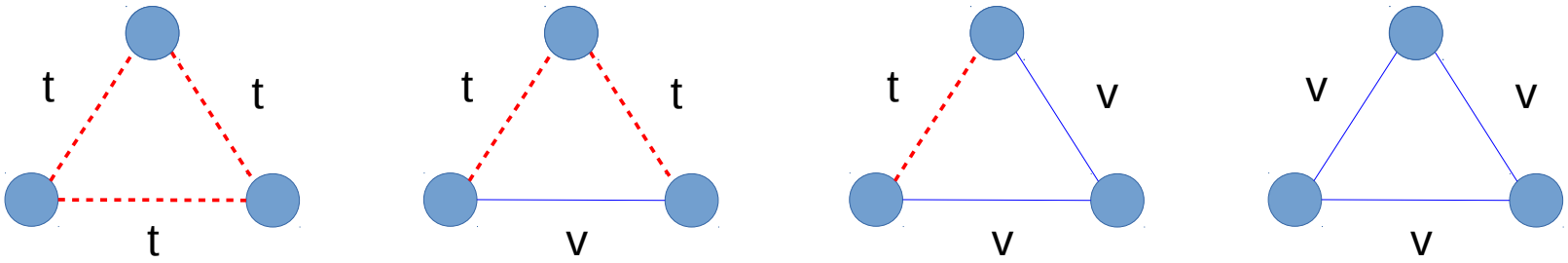
Star Wars



Ghostbusters



Network features



+3.73 -6.48 -1.05 +1.23

Homogeneity is preferred, but some forms of heterogenity are better than others.

Impact and next steps

- Our address term lexicons were used in a PNAS paper on racial disparities in police language during traffic stops (Voigt et al., 2017).
- Ongoing multidisciplinary effort on creation of social meaning in language (Pavalanathan et al., 2017; Kiesling et al., 2018).

Three short pieces

- **Exploring the construction of social meaning in networks**

  (Krishnan & Eisenstein, 2015; Pavalanathan et al., 2017)

- Operationalizing sociocultural influence from spatiotemporal cascades

  (Eisenstein et al., 2014; Goel et al., 2016; Soni & Eisenstein, 2018)

- Measuring the causal impact of efforts to combat hate speech

  (Chandrasekharan et al., 2018)

Three short pieces

- Exploring the construction of social meaning in networks

  (Krishnan & Eisenstein, 2015; Pavalanathan et al., 2017)

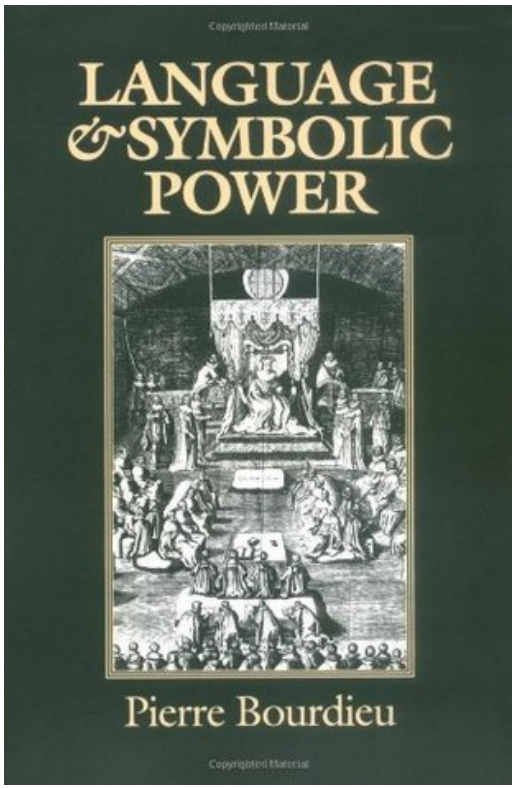
- **Operationalizing sociocultural influence from spatiotemporal cascades**

  (Eisenstein et al., 2014; Goel et al., 2016; Soni & Eisenstein, 2018)

- Measuring the causal impact of efforts to combat hate speech

  (Chandrasekharan et al., 2018)

Operationalizing sociocultural influence



- Bourdieu (1991): a key form of influence ("symbolic power") is the ability to control how others use language.
- Language change contains clues about who holds sociocultural influence.

Can we observe linguistic influence in real time?

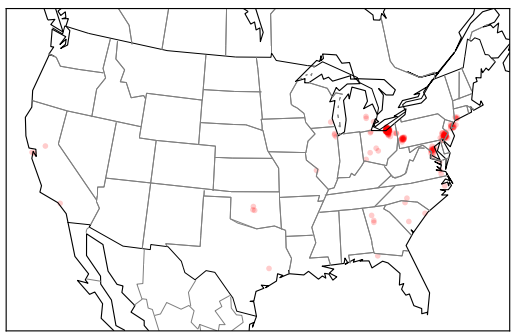
Can we observe linguistic influence in real time?

@name lmao! haahhaa ctfu!

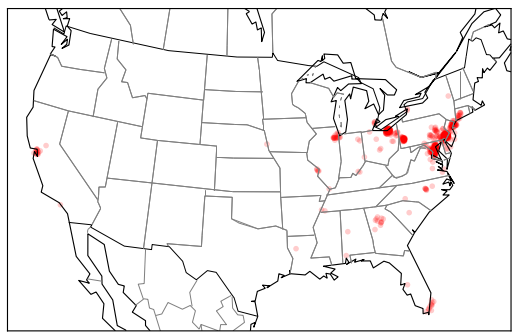
Can we observe linguistic influence in real time?

@name lmao! haahhaa ctfu!

2009



2010



2011

Twitter fads as social scientific evidence

- Propagating an innovation like ctfu requires:
  1. **Exposure**
  2. **Decision** to adopt

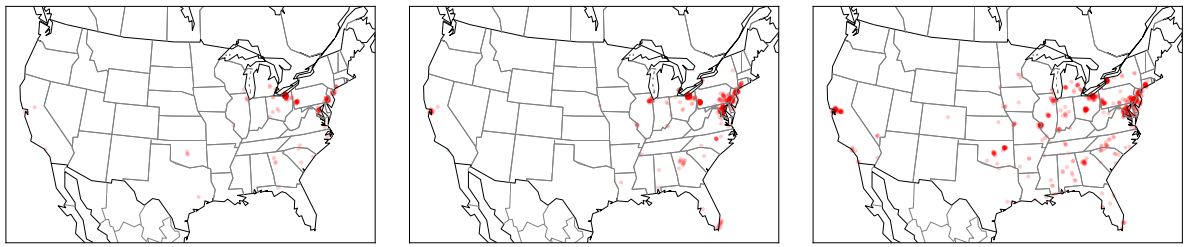
  (Rogers, 1962)

- By tracking the spread of these words, it is possible to reconstruct "deep networks" of social affinity and influence.

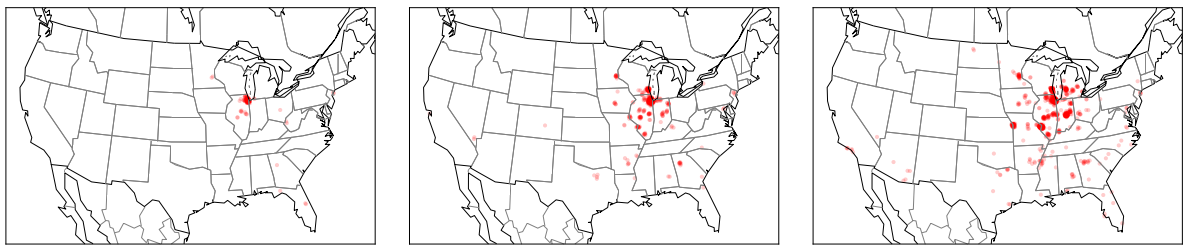
  (Eisenstein et al., 2014; Goel et al., 2016).

Thousands of words have changing frequencies.

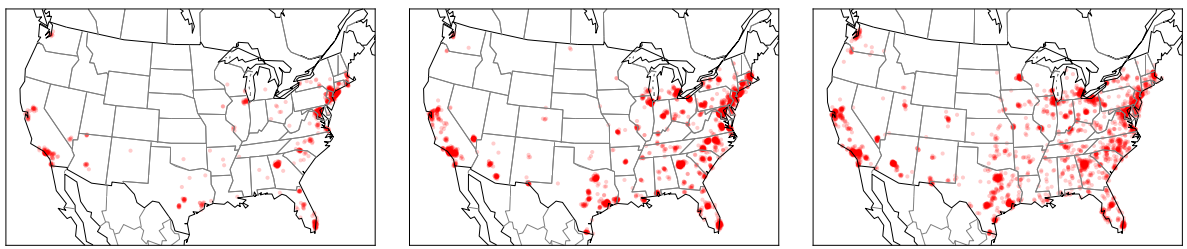
ctfu



lbvs



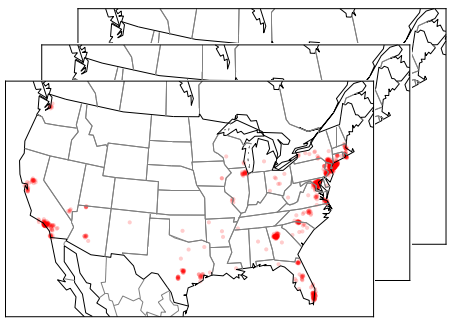
-__-



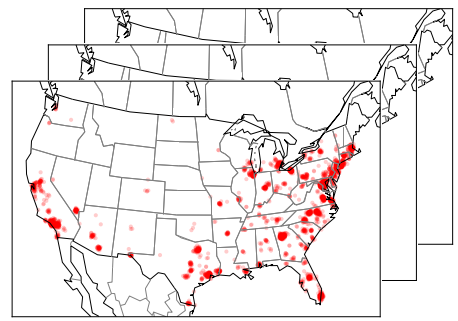
- Each spatiotemporal trajectory is idiosyncratic.
- What's the aggregate picture?

Language change as an autoregressive process

Raw data: counts for thousands of words, binned into 200 metro areas and 165 weeks.



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



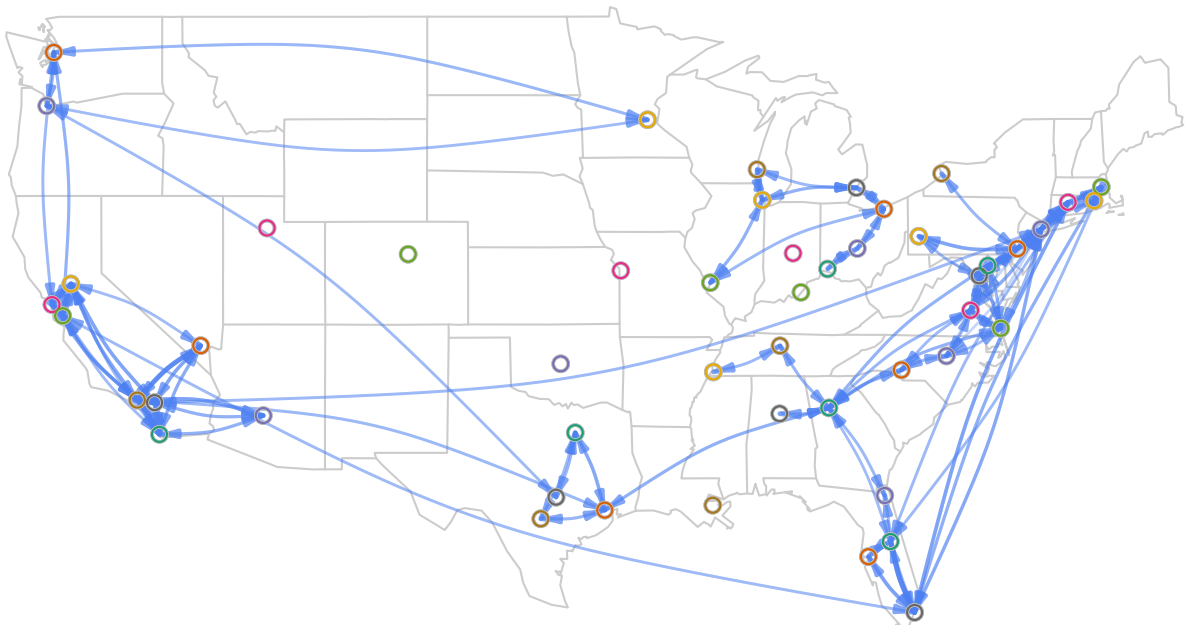
$\eta_3 \sim N(A\eta_2, \Sigma)$

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

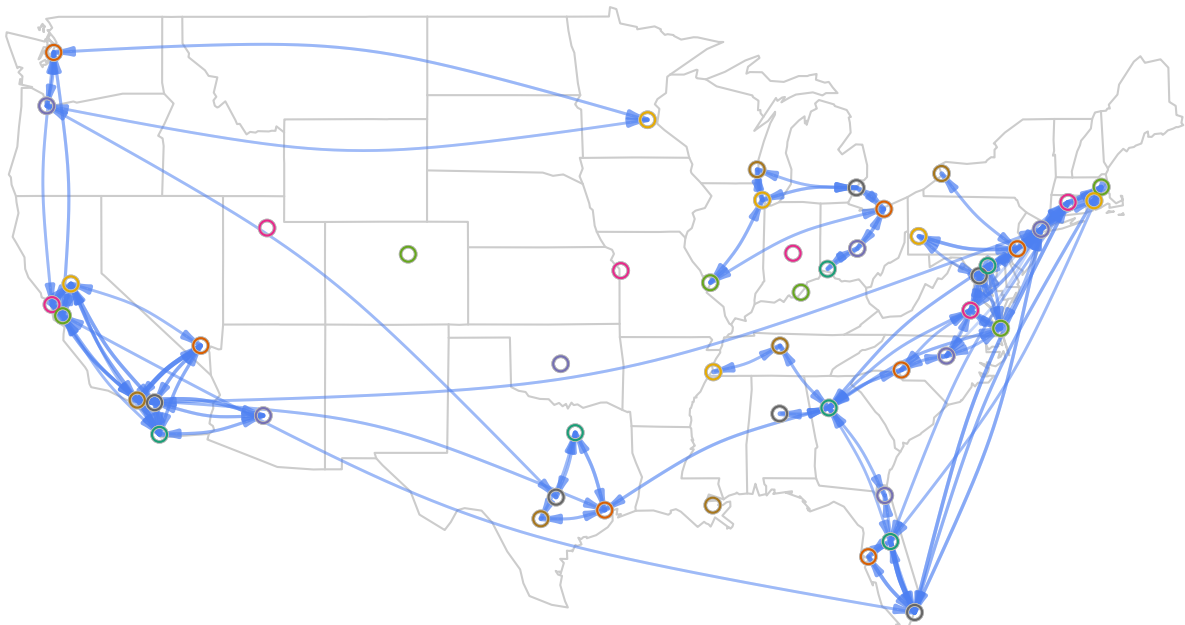
$$
\begin{aligned}
c_{\text{ctfu},2} &\sim \text{Binomial}(f(\eta_{\text{ctfu},2}), N_2) \\
c_{\text{hella},2} &\sim \text{Binomial}(f(\eta_{\text{hella},2}), N_2) \\
&\dots
\end{aligned}
$$

The dynamics matrix A encodes city-to-city linguistic influence (Eisenstein et al., 2014).

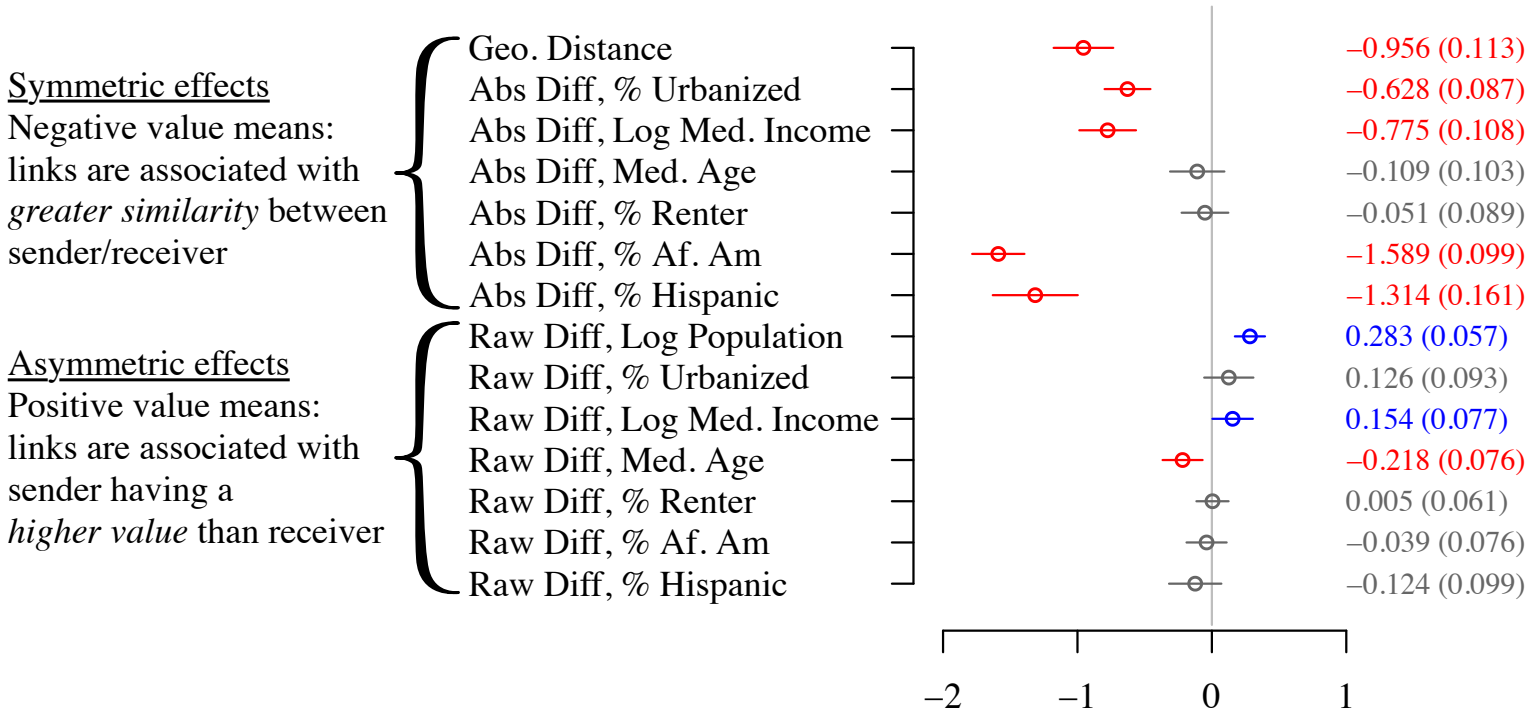
Aggregated city-to-city influence



Aggregated city-to-city influence



Is geography the whole story?



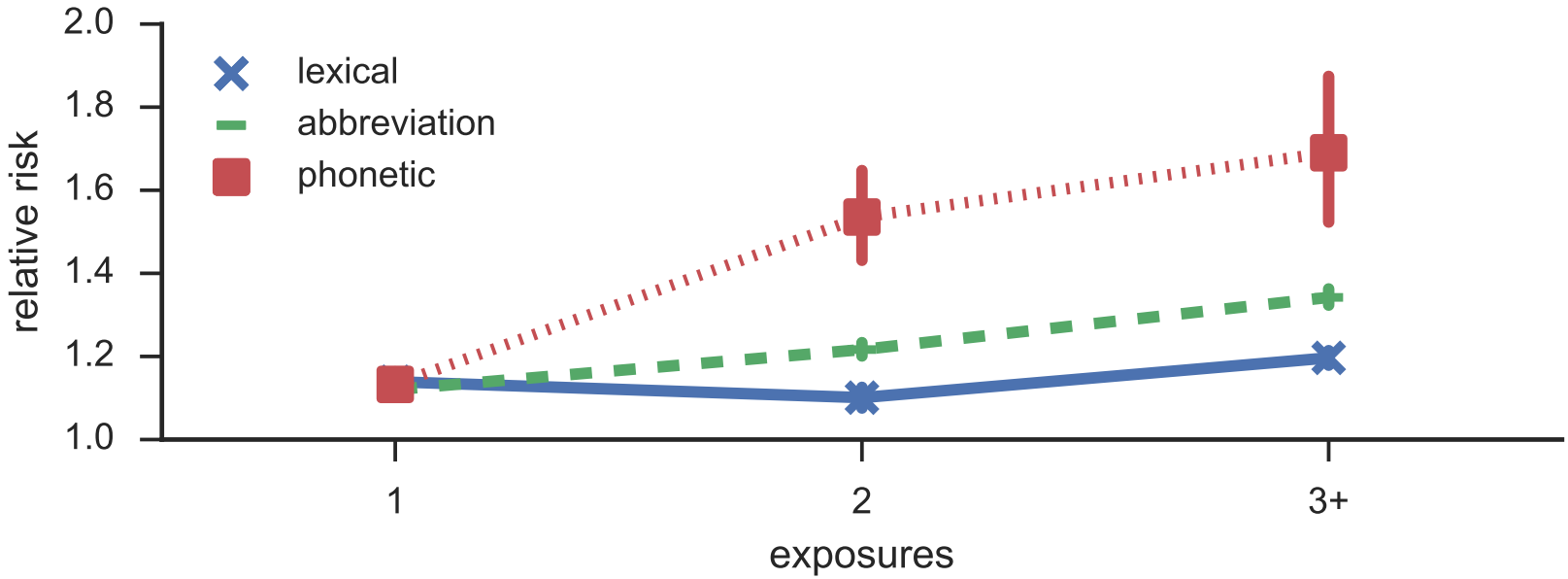
- Assortativity by race (of cities!) even more important than geography.
- Asymmetric effects are weaker, but bigger, younger metros tend to lead.

Person-to-person influence: whodunnit?

- Person i first uses a new linguistic feature at time t. **Who is responsible?**
- Suspect j should have the following properties:
  - Used the same feature at a time $t' < t$
  - Likely to be observed by i
- A random sample does not suffice!
  - I collaborated with MSR to obtain complete public records for 4.4M Twitter users, and their social networks.

Does language change on the network?

Does language change on the network?

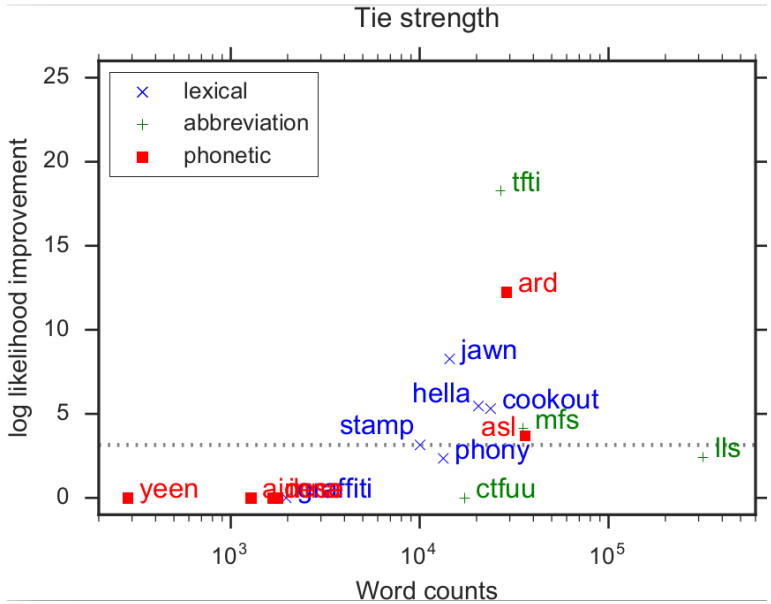


- Relative risk: likelihood of infection given exposure, normalized against rate in randomly-rewired network.
- For phonetic variables, risk increases with multiple exposures, a characteristic of complex contagion.

Which exposures are most influential?

Which exposures are most influential?

- We use a Parametric Hawkes Process to model the competing effects of multiple factors on a cascade of events.
- Challenge: scaling up to four million nodes.
- More mutual friends $\rightarrow$ more linguistic influence.



Three short pieces

- Exploring the construction of social meaning in networks

  (Krishnan & Eisenstein, 2015; Pavalanathan et al., 2017)

- **Operationalizing sociocultural influence from spatiotemporal cascades**

  (Eisenstein et al., 2014; Goel et al., 2016; Soni & Eisenstein, 2018)

- Measuring the causal impact of efforts to combat hate speech

  (Chandrasekharan et al., 2018)

Three short pieces

- Exploring the construction of social meaning in networks

  (Krishnan & Eisenstein, 2015; Pavalanathan et al., 2017)

- Operationalizing sociocultural influence from spatiotemporal cascades

  (Eisenstein et al., 2014; Goel et al., 2016; Soni & Eisenstein, 2018)

- **Measuring the causal impact of efforts to combat hate speech**

  (Chandrasekharan et al., 2018)

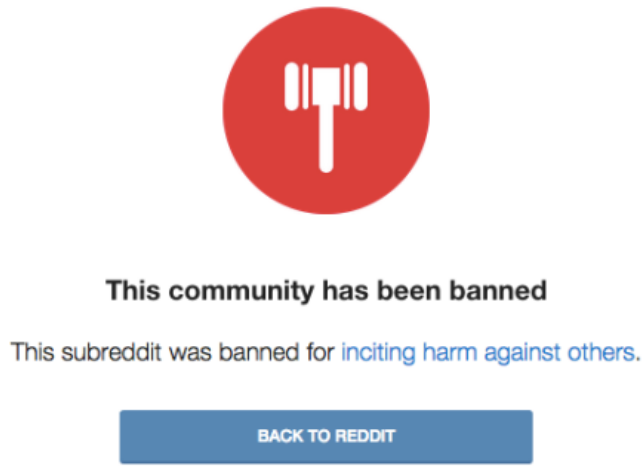
Hate speech on Reddit

What happens when forums for hate speech are shut down?

- Do participants export hate speech elsewhere?
- Or does the elimination of the "echo chamber" reduce hate speech overall?

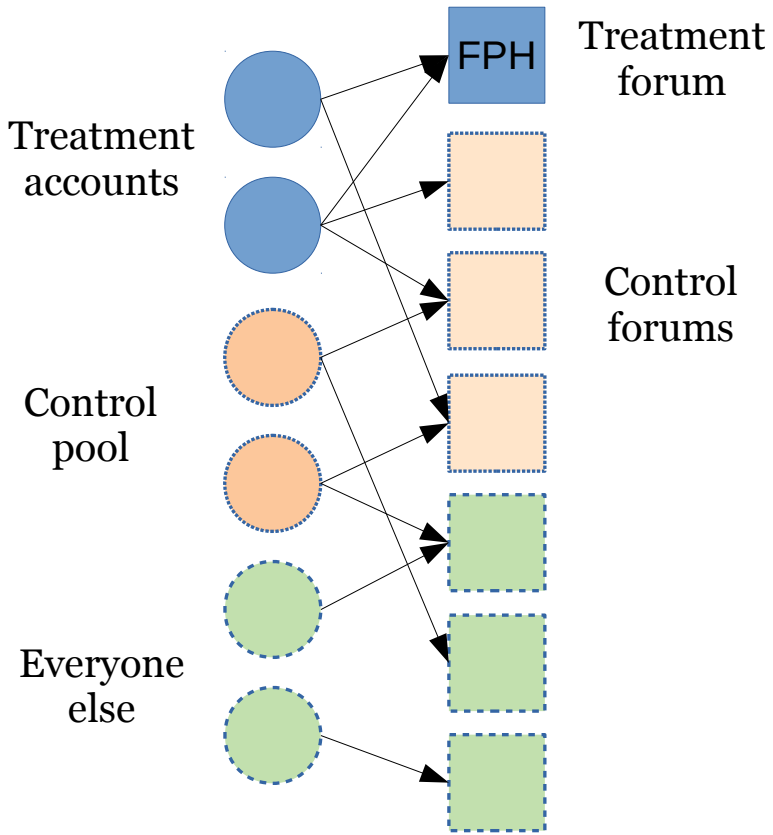
A natural experiment

- In 2015, Reddit closed several forums for violations of its anti-harassment policy.
- This enables a **natural experiment** on the effectiveness of this intervention (Chandrasekharan et al., 2018).



Causal inference design

- **Treatment group**: user accounts that post in the forums that were banned
- **Control forums**: other forums where the treatment group posts
- **Control pool**: other accounts who post in the control forums
- **Control group**: user accounts selected by Mahalanobis Distance Matching in the control pool

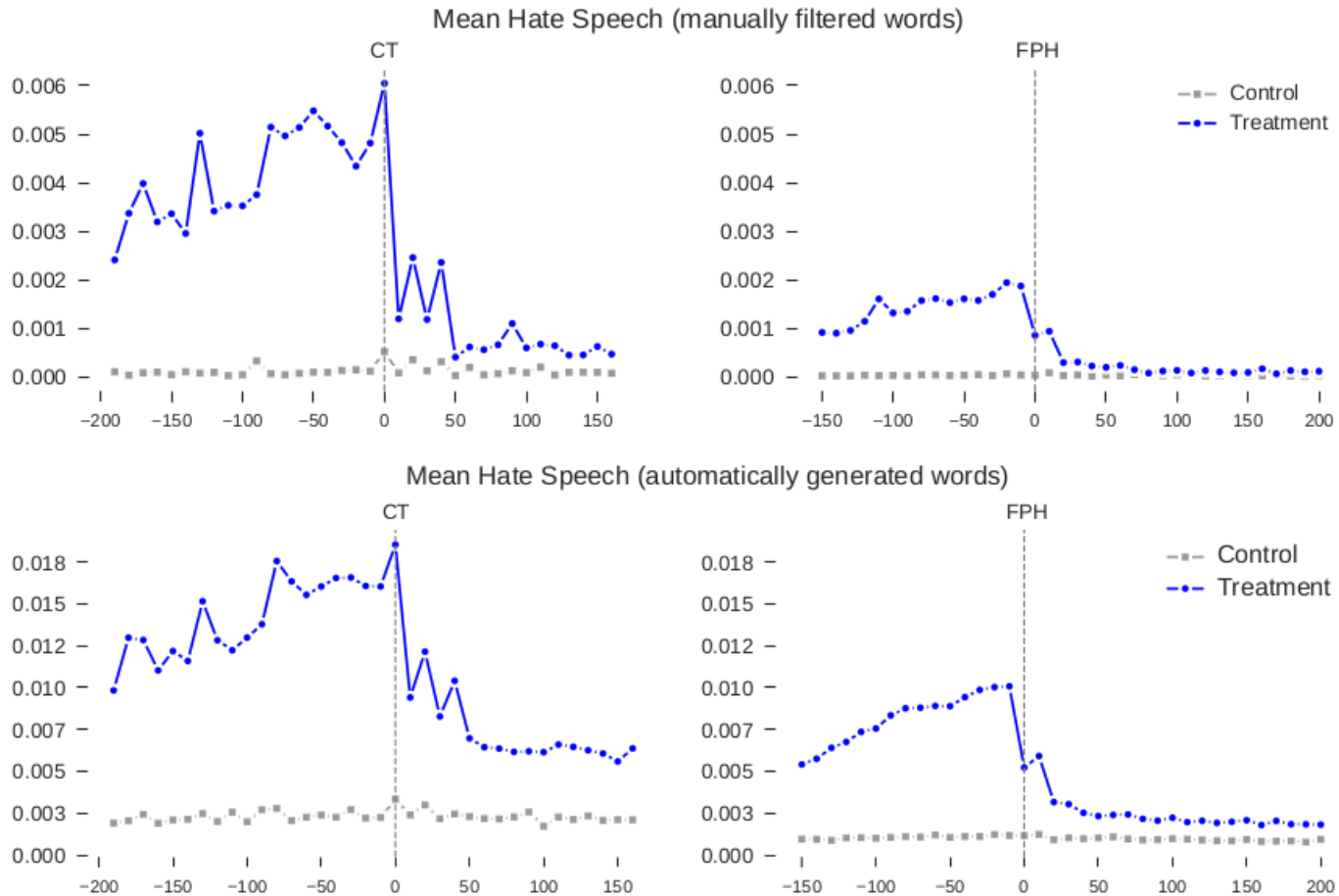


Measuring hate speech

1. Identify words that are unusually frequent in each forum, using SAGE (Eisenstein et al., 2011).
2. Examine the top 100, manually remove words that are not intrinsically linked to hate speech (EU Court of Human Rights definition)
   - the forum itself: fph, ct
   - the act of posting offensive content: $\mathrm{shitposting}$, $\mathrm{shitlord}$
   - words often used in non-hate speech contexts: IQ, $\mathrm{welfare}$, $\mathrm{cellulite}$

   We kept 20% of the original lexicon, $\kappa \approx .88$

Causal effect on hate speech



Aftermath

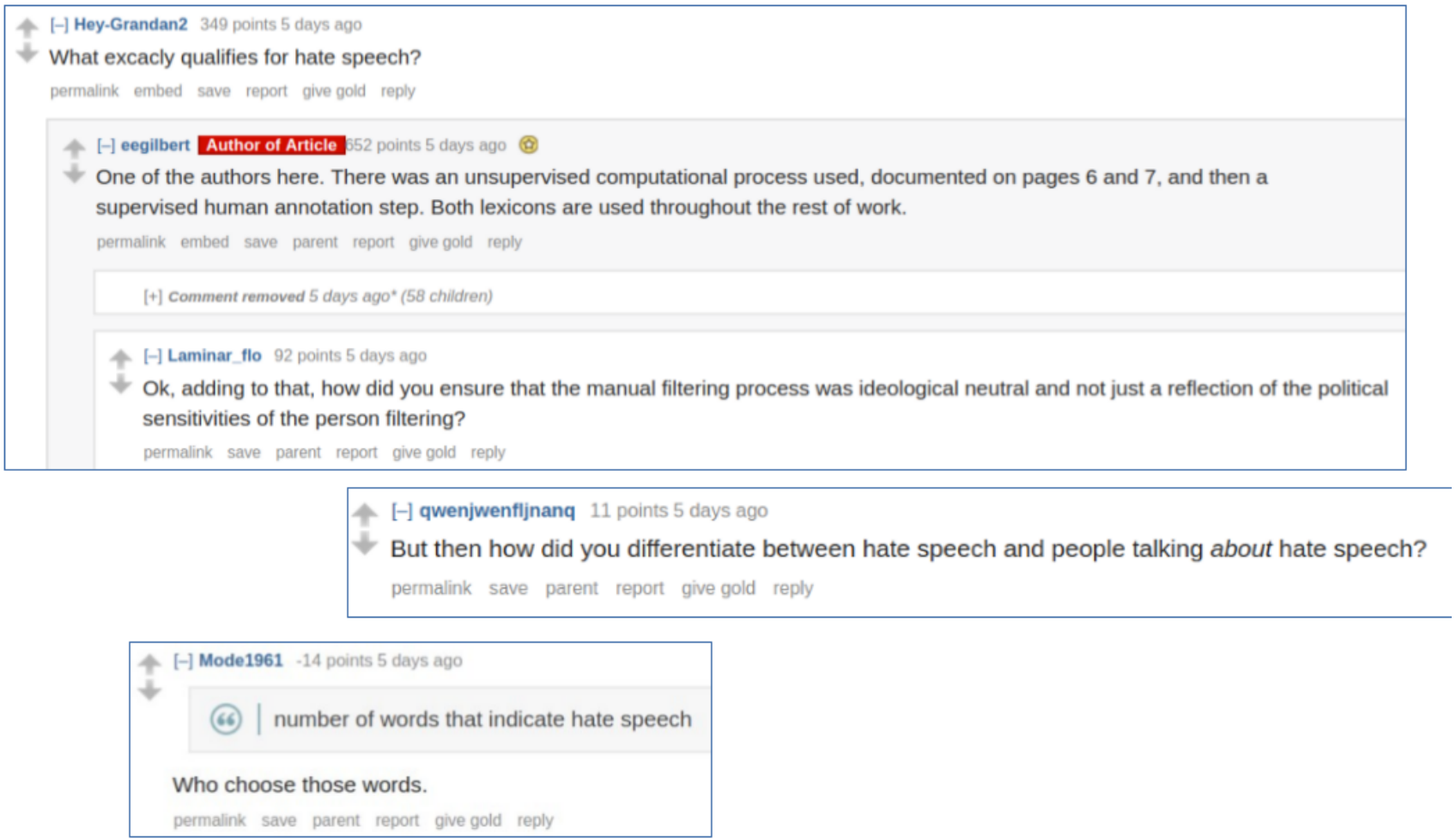
Reddit's bans of r/coontown and r/fatpeoplehate worked--many accounts of frequent posters on those subs were abandoned, and those who stayed reduced their use of hate speech ▸ comp.social.gatech.edu

5 months ago by asbruckman

Professor | Interactive Computing x2

6649 comments share save hide report

Aftermath



[–] **Hey-Grandan2** 349 points 5 days ago

What excacly qualifies for hate speech?

permalink embed save report give gold reply

[–] **eegilbert** **Author of Article** 652 points 5 days ago

One of the authors here. There was an unsupervised computational process used, documented on pages 6 and 7, and then a supervised human annotation step. Both lexicons are used throughout the rest of work.

permalink embed save parent report give gold reply

[+] *Comment removed 5 days ago* (58 children)*

[–] **Laminar_flo** 92 points 5 days ago

Ok, adding to that, how did you ensure that the manual filtering process was ideological neutral and not just a reflection of the political sensitivities of the person filtering?

permalink save parent report give gold reply

[–] **qwenjwenfljnanq** 11 points 5 days ago

But then how did you differentiate between hate speech and people talking *about* hate speech?

permalink save parent report give gold reply

[–] **Mode1961** -14 points 5 days ago

> number of words that indicate hate speech

Who choose those words.

permalink save parent report give gold reply

Aftermath

U.S.

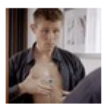
Reddit Bans Nazi Groups and Others in Crackdown on Violent Content

By **CHRISTINE HAUSER** OCT. 26, 2017

Steve Huffman, a co-founder and chief executive of Reddit, in 2016. The company has started to implement a new policy to remove content that glorifies and incites violence from its site. David Paul Morris/Bloomberg

RELATED COVERAGE

ON TECHNOLOGY
How Hate Groups Forced Online Platforms to Reveal Their True Nature AUG. 21, 2017



Opinion | Op-Ed Contributor
My Time Undercover With the Alt-Right
SEPT. 27, 2017

THE SHIFT
This Was the Alt-Right's Favorite Chat App. Then Came Charlottesville. AUG. 15, 2017

THE SHIFT
Reddit Limits Noxious Content by Giving Trolls Fewer Places to Gather SEPT. 25, 2017

(Why) did it work for Reddit?

- Reddit's federated structure delegates norm enforcement to moderators.

  It would be hard for Facebook and Twitter to target hate speech *communities* in the same way

- Some users went to alternative sites like Voat.

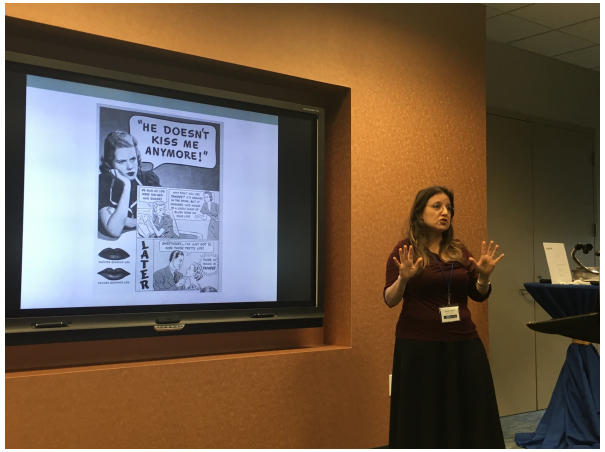
  Still a win for Reddit?

- Our algorithms detect only specific subsets of hate speech.

  Did hate speech shift to a form that is harder to detect?

Building a community

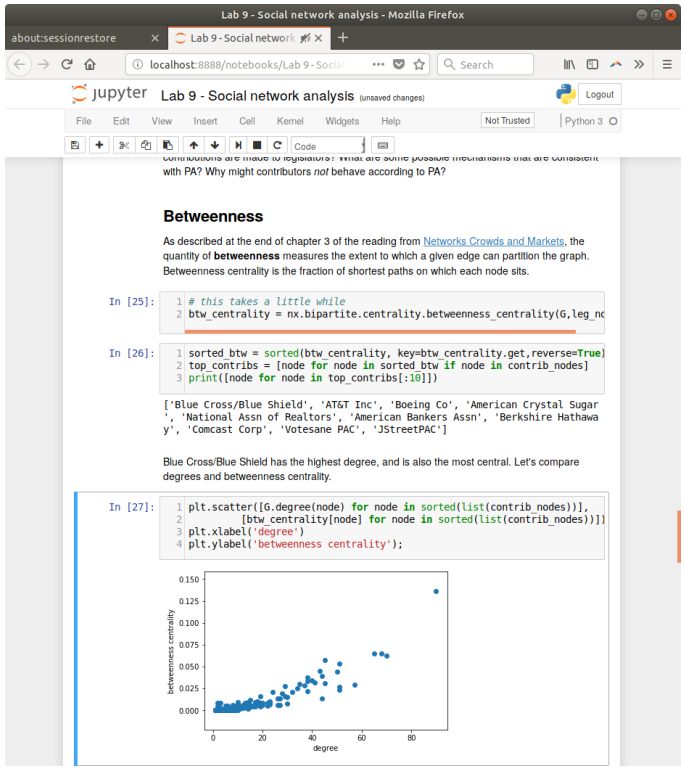
- NSF-funded doctoral consortium at EMNLP 2016
- NSF-funded workshop at ACL 2014
- Atlanta Computational Social Science Workshops at Emory, Georgia Tech, and Georgia State
- Panel on computational sociolinguistics at AAAS



Teaching

- Created CS8803-CSS: **Computational Social Science**
- Redesigned CS4464 and CS6465: **Computational Journalism**
- Redesigned CS4650/7650: **Natural Language**

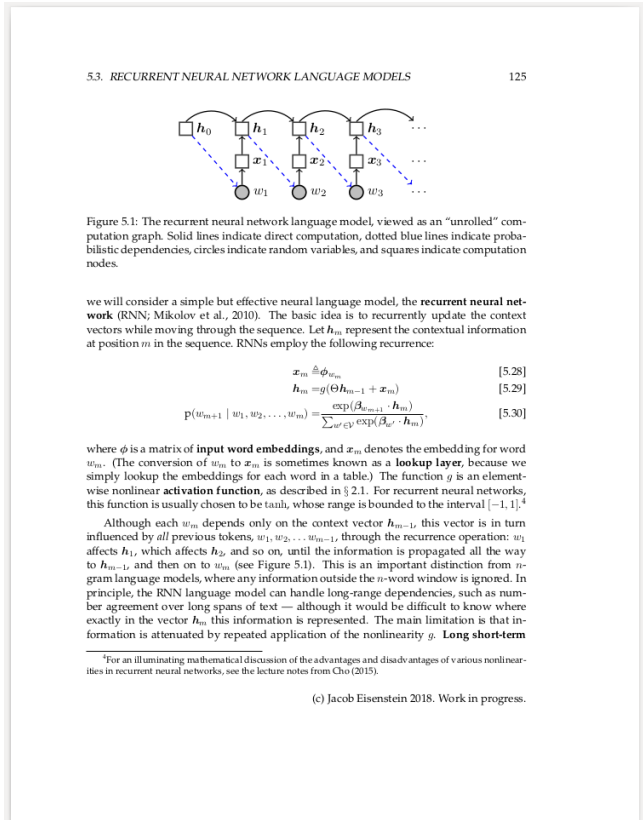
  Forthcoming textbook with MIT Press



Teaching

- Created CS8803-CSS: **Computational Social Science**
- Redesigned CS4464 and CS6465: **Computational Journalism**
- Redesigned CS4650/7650: **Natural Language**

  Forthcoming textbook with MIT Press



5.3. RECURRENT NEURAL NETWORK LANGUAGE MODELS 125

Figure 5.1: The recurrent neural network language model, viewed as an "unrolled" computation graph. Solid lines indicate direct computation, dotted blue lines indicate probabilistic dependencies, circles indicate random variables, and squares indicate computation nodes.

we will consider a simple but effective neural language model, the **recurrent neural network** (RNN; Mikolov et al., 2010). The basic idea is to recurrently update the context vectors while moving through the sequence. Let $\boldsymbol{h}_m$ represent the contextual information at position m in the sequence. RNNs employ the following recurrence:

$$\boldsymbol{x}_m \triangleq \boldsymbol{\phi}_{w_m} \quad [5.28]$$

$$\boldsymbol{h}_m = g(\boldsymbol{\Theta}\boldsymbol{h}_{m-1} + \boldsymbol{x}_m) \quad [5.29]$$

$$\mathrm{p}(w_{m+1} \mid w_1, w_2, \ldots, w_m) = \frac{\exp(\boldsymbol{\beta}_{w_{m+1}} \cdot \boldsymbol{h}_m)}{\sum_{w' \in \mathcal{V}} \exp(\boldsymbol{\beta}_{w'} \cdot \boldsymbol{h}_m)}, \quad [5.30]$$

where $\boldsymbol{\phi}$ is a matrix of **input word embeddings**, and $\boldsymbol{x}_m$ denotes the embedding for word w_m. (The conversion of w_m to $\boldsymbol{x}_m$ is sometimes known as a **lookup layer**, because we simply lookup the embeddings for each word in a table.) The function g is an element-wise nonlinear **activation function**, as described in § 2.1. For recurrent neural networks, this function is usually chosen to be $\tanh$, whose range is bounded to the interval $[-1, 1]$.[4]

Although each w_m depends only on the context vector $\boldsymbol{h}_{m-1}$, this vector is in turn influenced by *all* previous tokens, $w_1, w_2, \ldots w_{m-1}$, through the recurrence operation: w_1 affects $\boldsymbol{h}_1$, which affects $\boldsymbol{h}_2$, and so on, until the information is propagated all the way to $\boldsymbol{h}_{m-1}$, and then on to w_m (see Figure 5.1). This is an important distinction from n-gram language models, where any information outside the n-word window is ignored. In principle, the RNN language model can handle long-range dependencies, such as number agreement over long spans of text — although it would be difficult to know where exactly in the vector $\boldsymbol{h}_m$ this information is represented. The main limitation is that information is attenuated by repeated application of the nonlinearity g. **Long short-term**

[4] For an illuminating mathematical discussion of the advantages and disadvantages of various nonlinearities in recurrent neural networks, see the lecture notes from Cho (2015).

(c) Jacob Eisenstein 2018. Work in progress.

Summary

Computer and social science are inextricably linked:

- First wave CSS: large-scale instrumentation, crowd-sourcing, social network analysis
- Next wave CSS: artificial intelligence for semantic measurement of social phenomena
- Full circle: social scientific insights for better AI (Yang et al., 2016; Yang & Eisenstein, 2017).

Acknowledgments

- **Students**: Umashanthi Pavalanathan, Vinodh Krishnan, Yi Yang, Yangfeng Ji, Sandeep Soni, Ian Stewart, Yuval Pinter, Rahul Goel, Eshwar Chandrasekharan, Jim Fitzpatrick
- **Collaborators**: Ming-Wei Chang, Munmun De Choudhury, Eric Gilbert, Scott Kiesling, Lauren F. Klein, Dong Nguyen, Brendan O'Connor, Noah A. Smith, Eric P. Xing
- **Sponsors**: NSF, AFOSR, NIH, DTRA, NEH, Google

Summary

Computer and social science are inextricably linked:

- First wave CSS: large-scale instrumentation, crowd-sourcing, social network analysis
- Next wave CSS: artificial intelligence for semantic measurement of social phenomena
- Full circle: social scientific insights for better AI (Yang et al., 2016; Yang & Eisenstein, 2017).

References I

Bourdieu, P. (1991). *Language and symbolic power*. Harvard University Press.

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & Gilbert, E. (2018). You can't stay here: The effectiveness of Reddit's 2015 ban through the lens of hate speech. In *Proceedings of Computer-Supported Cooperative Work (CSCW)*.

Danescu-Niculescu-Mizil, C. & Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics*.

Eisenstein, J., Ahmed, A., & Xing, E. P. (2011). Sparse additive generative models of text. In *Proceedings of the International Conference on Machine Learning (ICML)*, (pp. 1041–1048).

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)*.

Gutmann, M. U. & Hyvärinen, A. (2012). Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *The Journal of Machine Learning Research*, *13*(1), 307–361.

Kiesling, S. F., Pavalanathan, U., Fitzpatrick, J., Han, X., & Eisenstein, J. (2018). Interactional stancetaking in online forums. *Computational Linguistics*, *(in review)*.

Krishnan, V. & Eisenstein, J. (2015). "You're Mr. Lebowski, I'm The Dude": Inducing address term formality in signed social networks. In *NAACL*.

Pavalanathan, U., Fitzpatrick, J., Kiesling, S. F., & Eisenstein, J. (2017). A multidimensional lexicon for interpersonal stancetaking. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Rogers, E. M. (1962). *Diffusion of innovations*. New York: The Free Press.

Soni, S. & Eisenstein, J. (2018). Predict to explain: Measuring social influence through classifier comparison. In *In review at KDD 2018*.

Voigt, R., Camp, N. P., Prabhakaran, V., Hamilton, W. L., Hetey, R. C., Griffiths, C. M., Jurgens, D., Jurafsky, D., & Eberhardt, J. L. (2017). Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, 201702413.

References II

West, R., Paskov, H., Leskovec, J., & Potts, C. (2014). Exploiting social network structure for person-to-person sentiment analysis. *Transactions of the Association for Computational Linguistics*, *2*, 297–310.

Yang, Y., Chang, M.-W., & Eisenstein, J. (2016). Toward socially-infused information extraction: Embedding authors, mentions, and entities. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Yang, Y. & Eisenstein, J. (2017). Overcoming language variation in sentiment analysis with social attention. *Transactions of the Association for Computational Linguistics (TACL)*, *5*.