Sociolinguistic Structure Induction

Jacob Eisenstein
@jacobbeisenstein

Georgia Institute of Technology

January 28, 2016
Linguistics and social network analysis both wrestle with understanding emergent structure.

- Linguistics: how does meaning arise from individual linguistic features?
- Social networks: how do macro-level social phenomena arise from individual dyads?

**Premise:** Jointly modeling linguistic and social structures will help us capture the emergence of social meaning in communication.
Overview

1. Language standards and innovation in social media networks
2. Linguistic signals of interpersonal relationships
3. Social communities for more robust NLP

Thanks to support from the National Science Foundation, the Air Force Office of Scientific Research, and the National Institutes for Health.
Language standards and innovation in social media networks

yinz

ard ("alright")

lbvs ("laughing but very serious")

ctfu ("cracking the fuck up")

(Eisenstein et al., 2010, 2014)
A new geographical diversity?

What explains these geographical differences in online writing?

- These are not just variables from speech!
- Are the users of these words aware that they are geographically specific?
- Will these differences eventually “wash out”?
- Modeling usage in context can shed light on the social function of linguistic variables.
Methods XV @MethodsXV · May 15

Our full programme will follow in couple of days! We’re very excited about it - so many great talks!

Broadcast
Methods XV @MethodsXV · Aug 11 · More
#methodsxv has officially opened pic.twitter.com/A4u2Zeuy8U

Hashtag-initial
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 $m$. 

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)
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 → less standard language

Local ties make non-standard language even more likely.
Local audience $\rightarrow$ less standard language

More mentions by users in same metro area

More mentions by users in other metro areas

Messages containing local variable

Messages not containing local variable
A new geographical diversity?

What explains these geographical differences in online writing?

- These are not just variables from speech!
- Are the users of these words aware that they are geographically specific?
- Will these differences eventually “wash out”?
- Modeling usage in context can shed light on the social function of linguistic variables.
Linguistic Signals of Social Relationships

You’re Mr. Lebowski, I’m the Dude: Inducing Address Term Formality in Signed Social Networks
I’m not Mr. Lebowski. You’re Mr. Lebowski. I’m the Dude.

- Jeff Bridges is The Dude
The Big Lebowski
Modeling formality of address

- What is the nature of the relationships between actors in a social network?

- Are there regular structures that emerge across signed networks?

- How does language reflect and reproduce social relationships?
Modeling formality of address

- What is the nature of the relationships between actors in a social network?
  
  We induce a clustering of edges by formality.

- Are there regular structures that emerge across signed networks?

- How does language reflect and reproduce social relationships?
Modeling formality of address

- What is the nature of the relationships between actors in a social network?
  
  We induce a clustering of edges by formality.

- Are there regular structures that emerge across signed networks?
  
  We estimate a prior distribution over signed network structures.

- How does language reflect and reproduce social relationships?
Modeling formality of address

- What is the nature of the relationships between actors in a social network?
  We induce a clustering of edges by formality.

- Are there regular structures that emerge across signed networks?
  We estimate a prior distribution over signed network structures.

- How does language reflect and reproduce social relationships?
  We estimate a likelihood distribution over address terms given formality.
A generative model of networked content

- Undirected graph $G = \langle i, j \rangle, \ i < j$

- All edges have labels $y_{ij} \in Y$

- Content vectors $x_i \rightarrow j, x_j \leftarrow i \in N^V$, where $V$ is the size of the vocabulary

- Each edge label has two associated distributions, so that:
  
  $x_i \rightarrow j | y_{ij} \sim \text{Multinomial}(\theta \rightarrow y_{ij})$

  $x_i \leftarrow j | y_{ij} \sim \text{Multinomial}(\theta \leftarrow y_{ij})$.

- We add the constraint $\theta \rightarrow y_{ij} = \theta \leftarrow y_{ij} = \theta_{y_{ij}}$; without it, we can generalize to directed signs.
A generative model of networked content

- Undirected graph $G = \langle i, j \rangle$, $i < j$
- All edges have labels $y_{ij} \in \mathcal{Y}$
A generative model of networked content

- Undirected graph $G = \langle i, j \rangle, \ i < j$
- All edges have labels $y_{ij} \in \mathcal{Y}$
- Content vectors $x_{i \rightarrow j}, x_{j \leftarrow i} \in \mathbb{N}^V$, where $V$ is the size of the vocabulary
A generative model of networked content

- Undirected graph $G = \langle i, j \rangle, \ i < j$
- All edges have labels $y_{ij} \in \mathcal{Y}$
- Content vectors $x_{i \rightarrow j}, x_{j \leftarrow i} \in \mathbb{N}^V$, where $V$ is the size of the vocabulary
- Each edge label has two associated distributions, so that:

  $x_{i \rightarrow j} \mid y_{ij} \sim \text{Multinomial}(\theta_{y_{ij}})$
  $x_{i \leftarrow j} \mid y_{ij} \sim \text{Multinomial}(\theta_{y_{ij}})$. 

We add the constraint $\theta_{y_{ij}} = \theta_{y_{ij}} = \theta_{y_{ij}}$; without it, we can generalize to directed signs.
A generative model of networked content

- Undirected graph $G = \langle i, j \rangle, \ i < j$
- All edges have labels $y_{ij} \in \mathcal{Y}$
- Content vectors $x_{i \rightarrow j}, x_{j \leftarrow i} \in \mathbb{N}^V$, where $V$ is the size of the vocabulary
- Each edge label has two associated distributions, so that:

$$x_{i \rightarrow j} \mid y_{ij} \sim \text{Multinomial}(\theta_{y_{ij}}^{\rightarrow})$$
$$x_{i \leftarrow j} \mid y_{ij} \sim \text{Multinomial}(\theta_{y_{ij}}^{\leftarrow}).$$

- We add the constraint $\theta_{y_{ij}}^{\rightarrow} = \theta_{y_{ij}}^{\leftarrow} = \theta_{y_{ij}}$; without it, we can generalize to directed signs.
Example

- At inference time, we observe $x$ but not $y$.
- Inferring $y$ gives a labeling over edges.
- Estimating $\theta$ gives the distribution over addresses for each edge type.

$y_{i2} = v$

$x_{i2} = \{\text{'Mr Lebowski':3, 'sir':1}\}$

$x_{21} = \{\text{'sir': 1}\}$
Adding social theory

- So far, this is just a mixture model over dyads.
- But social theory may tell us that not all label configurations are equally likely.
- Ex: **structural balance theory** describes networks of friend/enemy links, where signed triads may be stable or unstable:

<table>
<thead>
<tr>
<th>Strong structural balance</th>
<th>Stable</th>
<th>Unstable</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Stable Strong Structural Balance" /></td>
<td><img src="image2" alt="Unstable Strong Structural Balance" /></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Weak structural balance</th>
<th>Stable</th>
<th>Unstable</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image3" alt="Stable Weak Structural Balance" /></td>
<td><img src="image4" alt="Unstable Weak Structural Balance" /></td>
<td></td>
</tr>
</tbody>
</table>
Social theories with unknown parameters

- But what if the magnitude, and even the direction of the effect of each triad type is a priori unknown?

- We assume a triadic form, but make no assumptions about the specifics.
Social theory in a prior distribution

Assume the prior factors over dyads and triads.

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

where,

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

Assume the prior factors over dyads and triads.

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

where,

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

Assume the prior factors over dyads and triads.

\[ P(y; G, \eta, \beta) = \frac{1}{Z(\eta, \beta; G)} \times \exp \sum_{\langle i, j \rangle \in G} \eta^T f(y_{ij}, i, j, G) \]

\[ \times \exp \sum_{\langle i, j, k \rangle \in \mathcal{T}(G)} \beta_{y_{ij}, y_{jk}, y_{ik}}, \]

where,

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

Assume the prior factors over dyads and triads.

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

where,

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

Assume the prior factors over dyads and triads.

\[ P(y; G, \eta, \beta) = \frac{1}{Z(\eta, \beta; G)} \times \exp \sum_{\langle i,j \rangle \in G} \eta^\top f(y_{ij}, i, j, G) \]

\[ \times \exp \sum_{\langle i,j,k \rangle \in T(G)} \beta_{y_{ij},y_{jk},y_{ik}}, \]

where,

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

\[ P(y, x \mid G; \Theta, \beta, \eta) = P(x \mid y; \Theta)P(y \mid G; \beta, \eta) \]

- The likelihood factors across dyads;
- The prior factors across dyads and triads.
Complete model specification

\[ P(y, x \mid G; \Theta, \beta, \eta) = P(x \mid y; \Theta)P(y \mid G; \beta, \eta) \]

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

\[ P(y, x \mid G; \Theta, \beta, \eta) = P(x \mid y; \Theta)P(y \mid G; \beta, \eta) \]

- The likelihood factors across dyads;
- The \textbf{prior} factors across dyads and triads.
Complete model specification

\[ P(y, x \mid G; \Theta, \beta, \eta) = P(x \mid y; \Theta)P(y \mid G; \beta, \eta) \]

Bayesian inference answers several questions:
1. What is the label (formality) of each dyad?
2. How is formality expressed in language?
3. Are there structural regularities across networks?
Complete model specification

\[ P(y, x \mid G; \Theta, \beta, \eta) = P(x \mid y; \Theta)P(y \mid G; \beta, \eta) \]

Bayesian inference answers several questions:
1. What is the label (formality) of each dyad?
2. How is formality expressed in language?
3. Are there structural regularities across networks?
Complete model specification

\[ P(y, x \mid G; \Theta, \beta, \eta) = P(x \mid y; \Theta) P(y \mid G; \beta, \eta) \]

Bayesian inference answers several questions:
1. What is the label (formality) of each dyad?
2. How is formality expressed in language?
3. Are there structural regularities across networks?
Intractability of Inference

- The normalizing constant $z$ requires summing across all labelings.
  
  The number of labelings is $\mathcal{O}(\#|\mathcal{Y}|^N)$.

- West et al (2014) show that optimizing an objective over dyads and triads is NP-hard.
  
  Even exact posterior decoding of $\mathbf{y}$ is not tractable, given point estimates for parameters $\theta$, $\eta$, and $\beta$.

- We therefore apply a variational approximation.
  
  - We make point estimates of the parameters $\theta$, $\eta$, and $\beta$ (empirical Bayes).
  - We represent uncertainty over $\mathbf{y}$ through a factorized distribution $Q(\mathbf{y}) = \prod_{\langle i,j \rangle \in G} q_{ij}(y_{ij})$. 

Let’s do it!

- The Cornell Movie Dialogue corpus offers 300K conversational turns between 10K dyads, in 617 movies.
- All we need are the address terms...
- But no such resource appears to exist!

\[ y_{12} = v \]
\[ x_{12} = \{'Mr Lebowski':3, 'sir':1\} \]
\[ x_{21} = \{'sir': 1\} \]
Forms of address

Names such as Barack, Barack Hussein Obama.

Titles such as Ms., Dr., Private, Reverend. Titles can be used for address either by preceding a name (e.g., Colonel Kurtz), or in isolation (e.g., Yes, Colonel.).

Placeholder names such as dude, bro, brother, sweetie, cousin, and asshole. These terms can be used for address only in isolation.
Subtasks

- Build a vocabulary of titles.
- Build a vocabulary of placeholder names.
- Distinguish address tokens:
  \[\text{His/O name/O is/O Lebowski/O } ?/O \]
  \[\text{That’s/O your/O name/O, Dude/ADDR}\]

- There is surprisingly little prior work on these problems, including lexicons.
Automatic address annotations

Text: You’re Mr. Lebowski .
POS: PRP VBP NNP NNP
Address: O O O B-ADDR L-ADDR O

1. Look for character names (mined from rotten tomatoes).
2. Identify NNP tag sequences including those names.
3. Automatically label those sequences as entity spans.
Automatic address annotations

Text: You’re Mr. Lebowski.
POS: PRP VBP NNP NNP .
Address: O O O B-ADDR L-ADDR O

1. Look for character names (mined from rotten tomatoes).

2. Identify NNP tag sequences including those names.

3. Automatically label those sequences as entity spans.
Automatic address annotations

Text: You’re Mr. Lebowski.
POS: PRP VBP NNP NNP
Address: O O O B-ADDR L-ADDR O

1. Look for character names (mined from rotten tomatoes).
2. Identify NNP tag sequences including those names.
3. Automatically label those sequences as entity spans.
Automatic address annotations

Text: You ’re Mr. Lebowski .
POS: PRP VBP NNP NNP .
Address: O O O B-ADDR L-ADDR O

1. Look for character names (mined from rotten tomatoes).
2. Identify NNP tag sequences including those names.
3. Automatically label those sequences as entity spans.
## Address tagger: features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lexical</strong></td>
<td>The word to be tagged, and its two predecessors and successors, $w_{i-2:i+2}$</td>
</tr>
<tr>
<td><strong>POS</strong></td>
<td>The part-of-speech of the token to be tagged, and its neighbors.</td>
</tr>
<tr>
<td><strong>Case</strong></td>
<td>The case of the word to be tagged, and its neighbors.</td>
</tr>
<tr>
<td><strong>Constituency parse</strong></td>
<td>First non-NNP ancestor node of the word $w_i$ in the constituent parse tree, and all leaf node siblings in the tree.</td>
</tr>
<tr>
<td><strong>Dependency</strong></td>
<td>All dependency relations involving $w_i$.</td>
</tr>
<tr>
<td><strong>Location</strong></td>
<td>Distance of $w_i$ from the start and the end of the sentence or turn.</td>
</tr>
<tr>
<td><strong>Punctuation</strong></td>
<td>All punctuation symbols occurring before and after $w_i$.</td>
</tr>
<tr>
<td><strong>Second person pronoun</strong></td>
<td>All forms of the second person pronoun within the sentence.</td>
</tr>
</tbody>
</table>
Address tagger: accuracy

<table>
<thead>
<tr>
<th>Class</th>
<th>F-measure</th>
<th>Total Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>I-ADDR</td>
<td>0.58</td>
<td>53</td>
</tr>
<tr>
<td>B-ADDR</td>
<td>0.800</td>
<td>483</td>
</tr>
<tr>
<td>U-ADDR</td>
<td>0.987</td>
<td>1864</td>
</tr>
<tr>
<td>L-ADDR</td>
<td>0.813</td>
<td>535</td>
</tr>
<tr>
<td>O-ADDR</td>
<td>0.993</td>
<td>35975</td>
</tr>
</tbody>
</table>
LEXICON INDUCTION: TITLES

- Run the tagger, find terms that frequently appear at the beginning of address spans containing the character’s name.
- We then manually filter out 17 of 34 candidates, obtaining:

  agent, aunt, captain, colonel, commander, cousin, deputy, detective, dr, herr, inspector, judge, lord, master, mayor, miss, mister, miz, monsieur, mr, mrs, ms, professor, queen, reverend, sergeant, uncle
Lexicon induction: placeholder names

► Remove the current-word feature from the tagger model, then find terms that are frequently tagged as the unique element in an address span.
► After manually filter out 41 of 96 candidates, we obtain:

asshole, babe, baby, boss, boy, bro, bud, buddy, cocksucker, convict, cousin, cowboy, cunt, dad, darling, dear, detective, doll, dude, dummy, father, fella, gal, ho, hon, honey, kid, lad, lady, lover, ma, madam, madame, man, mate, mister, mon, moron, motherfucker, pal, papa, partner, peanut, pet, pilgrim, pop, president, punk, shithead, sir, sire, son, sonny, sport, sucker, sugar, sweetheart, sweetie, tiger
Feature vector construction

Content features

▶ Addressee name, including any title in lexicon
  You’re Mr. Lebowski → Mr. Last Name

▶ Any element in the placeholder name lexicon, if tagged as the unique element in address span
  Thanks, dude → DUDE

Dyad feature: Adamic-Adar metric (normalized mutual friends) for each dyad
# Model comparison

<table>
<thead>
<tr>
<th>Text</th>
<th>Dyad Feature</th>
<th>Signed triads</th>
<th>Predictive Log-likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M2</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>M3</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>M4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
## Model comparison

<table>
<thead>
<tr>
<th>Text</th>
<th>Dyad Feature</th>
<th>Signed triads</th>
<th>Predictive Log-likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>✓</td>
<td></td>
<td>-2133.28</td>
</tr>
<tr>
<td>M2</td>
<td>✓</td>
<td>✓</td>
<td>-2018.21</td>
</tr>
<tr>
<td>M3</td>
<td>✓</td>
<td>✓</td>
<td>-1884.02</td>
</tr>
<tr>
<td>M4</td>
<td>✓</td>
<td>✓</td>
<td>-1582.43</td>
</tr>
</tbody>
</table>

Predictive likelihood is evaluated on held-out address terms for a 10% test fold.
Cluster coherence

<table>
<thead>
<tr>
<th>V-cluster</th>
<th>T-cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>sir</td>
<td>FIRSTNAME</td>
</tr>
<tr>
<td>mr + LASTNAME</td>
<td>man</td>
</tr>
<tr>
<td>mr + FIRSTNAME</td>
<td>baby</td>
</tr>
<tr>
<td>mr</td>
<td>honey</td>
</tr>
<tr>
<td>miss + LASTNAME</td>
<td>darling</td>
</tr>
<tr>
<td>son</td>
<td>sweetheart</td>
</tr>
<tr>
<td>mister + FIRSTNAME</td>
<td>buddy</td>
</tr>
<tr>
<td>mrs</td>
<td>sweetie</td>
</tr>
</tbody>
</table>

On an intrusion task

- Raters found the intruder in 73% of cases for the full model (M4).
- ... versus 52% in the text-only model (M1).
Star Wars
Ghostbusters
South Park
Network features

\begin{align*}
\text{+3.73} & \quad \text{t} & \quad \text{t} & \quad \text{t} \\
\text{-6.48} & \quad \text{t} & \quad \text{v} & \quad \\
\text{-1.05} & \quad \text{t} & \quad \text{v} & \quad \text{v} \\
\text{+1.23} & \quad \text{v} & \quad \text{v} & \quad \text{v} \\
\end{align*}
Putting Things in Context: Community-specific Embedding Projections for Sentiment Analysis.
Yang & Eisenstein (2015)
Words mean different things to different people

“I would like to believe he’s sick rather than just mean and evil.”
Words mean different things to different people

“I would like to believe he’s sick rather than just mean and evil.”

“You could’ve been getting down to this sick beat.”

As author sets become increasingly diverse, variation is increasingly a problem for NLP.
Language diversity and social attributes

Language variation is everywhere. What can we do about it?


- What if we don’t have the author demographics?
Language diversity and social attributes

Language variation is everywhere. What can we do about it?

- What if we don’t have the author demographics?

Homophily: neighbors have similar properties.

- Demographics (Thelwall, 2009; Al Zamal et al., 2012)
- Language (Bryden et al., 2013)
Homophily to the rescue?

Social network structure can help us identify which meaning is likely, even in unlabeled instances.
The SemEval Social Network

- SemEval 2013-2015 is a standard benchmark for Twitter Sentiment Analysis (Nakov et al., 2013; Rosenthal et al., 2015).

- Follower network statistics:

<table>
<thead>
<tr>
<th></th>
<th># Author</th>
<th># Relations</th>
<th># Isolates</th>
</tr>
</thead>
<tbody>
<tr>
<td>original</td>
<td>14,087</td>
<td>40,110</td>
<td>3,633</td>
</tr>
<tr>
<td>expanded</td>
<td>17,417</td>
<td>1,050,369</td>
<td>689</td>
</tr>
</tbody>
</table>

We “densify” the social network by adding individuals who are followed by at least 100 SemEval authors.
From Homophily to Modularity

- Reasoning about lexical semantics in a social network of $10^4$ nodes is hard!
- Solution: summarize the network by clustering nodes into social network communities.
- **Modularity**: nodes in the same community tend to have similar properties.
### Linguistic Modularity in SemEval

<table>
<thead>
<tr>
<th>Network</th>
<th>Algorithm</th>
<th># Communities</th>
<th>$Q^{(L)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>original</td>
<td>Fast-Greedy</td>
<td>19</td>
<td>0.1549</td>
</tr>
<tr>
<td>expanded</td>
<td></td>
<td>5</td>
<td>0.2102</td>
</tr>
<tr>
<td>original</td>
<td>Multi-Level</td>
<td>19</td>
<td>0.1462</td>
</tr>
<tr>
<td>expanded</td>
<td></td>
<td>7</td>
<td>0.2526</td>
</tr>
</tbody>
</table>

**Table:** Linguistic modularities on the SemEval data

- Social network communities use similar words...
- But does meaning also cluster on a social network?
Nonlinear Subspace Embeddings (NLSE)

We build on NLSE (Astudillo et al., 2015), which learns a task-specific projection of word embeddings for sentiment analysis.

\[ h_j = \text{Sigmoid}(Sx_j), \quad (1) \]

where

- \( x_j \) is the pre-trained embedding of word \( j \)
- \( S \) is a discriminatively learned task-specific projection matrix
- \( h_j \) is the projected embedding
- To project multiple words, Astudillo et al. (2015) just add up the embeddings.
Adding Community Structure

For a node $i$ in community $c_i$, 

$$h_j = \text{Sigmoid} \left( \left[ S^{(0)} + S^{(c_i)} \right] w_j \right),$$  

We can also detect multiple community structures, and add separate projections for each.
## Results

<table>
<thead>
<tr>
<th>Communities</th>
<th>T2013</th>
<th>T2014</th>
<th>T2015</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prior work</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>72.0</td>
<td>72.7</td>
<td>65.2</td>
<td>70.0</td>
</tr>
<tr>
<td><strong>Our methods</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fast-Greedy</td>
<td>72.0</td>
<td>73.7</td>
<td>65.6</td>
<td>70.4</td>
</tr>
<tr>
<td>Multi-Level</td>
<td>72.1</td>
<td><strong>73.8</strong></td>
<td><strong>66.1</strong></td>
<td><strong>70.7</strong></td>
</tr>
</tbody>
</table>

**Table:** Average F1 score on the SemEval test sets. Results are marked with * if they are significant better than pooled NLSE at $p < 0.05$. Running NLSE separately in each community is substantially worse.
Overview

1. Language standards and innovation in social media networks

2. Linguistic signals of interpersonal relationships

3. Social communities for more robust NLP

Thanks to support from the National Science Foundation, the Air Force Office of Scientific Research, and the National Institutes for Health.
What’s next?

Beyond and inside the lexicon Tracing syntactic and orthographic markers of linguistic style and social dynamics.

From style to content What do your stylistic decisions reveal about what you believe?

From macro to micro Tracking the spread of language change through individual linguistic decisions.
Overview

1. Language standards and innovation in social media networks
2. Linguistic signals of interpersonal relationships
3. Social communities for more robust NLP

Thanks to support from the National Science Foundation, the Air Force Office of Scientific Research, and the National Institutes for Health.


Social networks in social media

Social media platforms offer a number of forms of metadata that capture social networks.

**Articulated network**  Explicitly-defined connections; undirected in Facebook, directed in Twitter.

**Behavioral network**  Inferred from conversational interactions, such as replies or mentions.  

(danah boyd, 2006)
Social networks on Twitter

- Twitter users often follow 1000s of other users.
- Mention networks are smaller, and arguably more socially meaningful.
- Twitter query rate limiting makes mention network much easier to obtain.

(Huberman et al., 2008)
Community-specific Meanings

Most distinct words

C1 brighter blessings celebrate shine brightest blessing gbu glowing celebrating hwaiting
C2 mistakes stfu ass shutup bitch #ohwell retard regret dgaf idgaf
C3 enjoyin #stillkidrauhl funny brighter happiest catchy sweeter jealous relaxing smartest