

# **Statistical Text Analysis for Social Science**

**Brendan T. O'Connor**

August 2014  
CMU-ML-14-101

Machine Learning Department  
School of Computer Science  
Carnegie Mellon University  
Pittsburgh, PA, USA

*Thesis Committee:*  
Noah A. Smith, chair  
Tom Mitchell  
Cosma Shalizi  
Gary King, Harvard University

*Submitted in partial fulfillment of the requirements  
for the Degree of Doctor of Philosophy*

This work was supported by NSF CAREER IIS0644225, NSF CAREER IIS1054319, NSF grant IIS1211277, NSF grant IIS1251131, DARPA grant N10AP20042, IARPA contract D12PC00347, an Alfred P. Sloan grant, Microsoft's support for Machine Learning graduate students, Google's support of the Worldly Knowledge project at Carnegie Mellon University, the Berkman Faculty Development Fund at Carnegie Mellon University, the Center for Applied Research in Technology at the Tepper School of Business, computing resources from the Open Source Data Cloud ([Grossman et al. \(2012\)](#), sponsored by the Open Cloud Consortium, the Gordon and Betty Moore Foundation, the NSF, and the University of Chicago) and computing resources from the Pittsburgh Supercomputing Center.

*Keywords:* computational social science, natural language processing, text mining, quantitative text analysis, machine learning, probabilistic graphical models, Bayesian statistics, exploratory data analysis, social media, opinion polling, sociolinguistics, event data, international relations.

---

# Abstract

---

What can text corpora tell us about society? How can automatic text analysis algorithms efficiently and reliably analyze the social processes revealed in language production?

This work develops statistical text analyses of dynamic social and news media datasets to extract indicators of underlying social phenomena, and to reveal how social factors guide linguistic production. This is illustrated through three case studies: first, examining whether sentiment expressed in social media can track opinion polls on economic and political topics (Chapter 3); second, analyzing how novel online slang terms can be very specific to geographic and demographic communities, and how these social factors affect their transmission over time (Chapters 4 and 5); and third, automatically extracting political events from news articles, to assist analyses of the interactions of international actors over time (Chapter 6).

We demonstrate a variety of computational, linguistic, and statistical tools that are employed for these analyses, and also contribute *MiTextExplorer*, an interactive system for exploratory analysis of text data against document covariates, whose design was informed by the experience of researching these and other similar works (Chapter 2). These case studies illustrate recurring themes toward developing text analysis as a social science methodology: computational and statistical complexity, and domain knowledge and linguistic assumptions.

---

# Acknowledgments

---

This work has been made possible only through rich collaborations and conversations with a wide variety of colleagues at both Carnegie Mellon (including the Noah’s ARK lab and the Machine Learning Department) and many other institutions. Thanks to everyone for all the support and learning that influenced this work, from people including but not limited to: Yisong Yue, Dani Yogatama, Tae Yano, Min Xu, Eric Xing, Matt Wytock, Hanna Wallach, Johan Ugander, Dustin Tingley, Ryan Tibshirani, Sam Thomson, Andrew Thomas, Eric Sun, Brandon Stewart, Rion Snow, Alex Smola, Noah Smith, Yanchuan Sim, Cosma Shalizi, Phil Schrodт, Nathan Schneider, Aaron Schein, Bryan Routledge, Molly Roberts, Philip Resnik, Joseph Reisinger, Dan Ramage, Patrick Perry, Olutobi Owoputi, John Myles White, Burt Monroe, Tom Mitchell, Brendan Meeder, Andre Martins, Cameron Marlow, Yucheng Low, Michael Love, Jayant Krishnamurthy, Michel Krieger, Lingpeng Kong, Gary King, Rob Kass, Venky Iyer, Muhammed Idris, Qirong Ho, Michael Heilman, Alex Hanna, Aria Haghghi, Forest Gregg, Joey Gonzalez, Kevin Gimpel, Matt Gardner, Alona Fyshe, Seth Flaxman, Jeffrey Flanigan, Jenny Finkel, Khalid El-Arini, Jacob Eisenstein, Chris Dyer, Jesse Dodge, Dipanjan Das, Justin Cranshaw, Fred Conrad, William Cohen, Shay Cohen, Jonathan Chang, Andrew Carlson, Dallas Card, Lukas Biewald, Justin Betteridge, John Beieler, David Bamman, Ramnath Balasubramanyan, Erin Baggott, Lars Backstrom, Waleed Ammar, and David Ahn. My apologies for any omissions. Thanks most of all to my family and friends for all their support and encouragement.

---

# Contents

---

<b>Contents</b>	<b>5</b>
<b>1 Text analysis for the social sciences</b>	<b>8</b>
1.1 Introduction . . . . .	8
1.2 Text analysis methods . . . . .	11
1.3 Thesis statement . . . . .	14
<b>2 <i>MiTextExplorer</i>: Interactive exploratory text and covariate analysis</b>	<b>15</b>
2.1 Introduction . . . . .	15
2.2 Motivation: Can we “just look” at statistical text data? . . . . .	16
2.3 MiTEXTEXPLORER: linked brushing for text and covariate correlations . . . . .	18
2.3.1 Covariate-driven queries . . . . .	19
2.3.2 Term selection and KWIC views . . . . .	20
2.3.3 Term-association queries . . . . .	21
2.3.4 Pinned terms . . . . .	22
2.4 Statistical term association measures . . . . .	23
2.5 Related work: Exploratory text analysis . . . . .	24
2.6 Future work . . . . .	26
<b>3 Text sentiment and opinion polling</b>	<b>27</b>
3.1 Introduction . . . . .	27
3.2 Data . . . . .	27
3.2.1 Twitter Corpus . . . . .	28
3.2.2 Public Opinion Polls . . . . .	28
3.3 Text Analysis . . . . .	29
3.3.1 Message Retrieval . . . . .	30
3.3.2 Opinion Estimation . . . . .	31
3.3.3 Comparison to Related Work . . . . .	32
3.4 Moving Average Aggregate Sentiment . . . . .	32
3.5 Correlation Analysis: Is text sentiment a leading indicator of polls? . . . . .	33
3.5.1 Obama 2009 Job Approval and 2008 Elections . . . . .	37
3.6 Measurement bias and inference goals . . . . .	38
3.7 Conclusion . . . . .	40

<b>4 Mixed-membership analysis of geographic and demographic lexical variation in social media</b>	<b>41</b>
4.1 Introduction . . . . .	41
4.2 Geographic topic model and lexical variation . . . . .	42
4.2.1 Synopsis . . . . .	42
4.2.2 Data . . . . .	43
4.2.3 Model . . . . .	44
4.2.4 Inference . . . . .	46
4.2.5 Implementation . . . . .	47
4.2.6 Evaluation . . . . .	48
4.2.7 Analysis . . . . .	51
4.2.8 Related Work . . . . .	53
4.3 Demographic lexical variation through U.S. Census data . . . . .	54
4.3.1 Introduction . . . . .	54
4.3.2 Data . . . . .	54
4.3.3 Model . . . . .	55
4.3.4 Analysis . . . . .	56
4.4 Conclusion and reflections on textual social data analysis . . . . .	58
<b>5 Social determinants of linguistic diffusion in social media</b>	<b>60</b>
5.1 Synopsis . . . . .	60
5.2 Introduction . . . . .	60
5.3 Materials and methods . . . . .	62
5.3.1 Modeling spatiotemporal lexical dynamics in social media data . . . . .	63
5.3.2 Constructing a network of linguistic diffusion . . . . .	70
5.3.3 Geographic and demographic correlates of linguistic diffusion . . . . .	71
5.4 Results . . . . .	73
5.5 Discussion . . . . .	75
5.6 Appendix: Data processing . . . . .	76
5.6.1 Messages . . . . .	77
5.6.2 Words . . . . .	78
<b>6 International Events from the News</b>	<b>85</b>
6.1 Synopsis . . . . .	85
6.2 Introduction . . . . .	85
6.3 Data . . . . .	86
6.4 Model . . . . .	87
6.4.1 Smoothing Frames Across Time . . . . .	89
6.5 Inference . . . . .	89
6.6 Experiments . . . . .	90
6.6.1 Lexical Scale Impurity . . . . .	90
6.6.2 Conflict Detection . . . . .	91
6.6.3 Results . . . . .	93
6.6.4 Case study . . . . .	94
6.7 Related Work . . . . .	95
6.7.1 Events Data in Political Science . . . . .	95
6.7.2 Events in Natural Language Processing . . . . .	96
6.8 Conclusion . . . . .	96

6.9	Appendix: Evaluation . . . . .	97
6.9.1	Type-level calculation of lexical scale impurity . . . . .	97
6.9.2	TABARI lexicon matching . . . . .	97
6.10	Appendix: Details on Inference . . . . .	98
6.10.1	Language Submodel $[z \mid \eta]$ . . . . .	99
6.10.2	Context Submodel $[\alpha, \beta \mid \eta]$ . . . . .	99
6.10.3	Logistic Normal $[\eta \mid z, \bar{\eta}]$ . . . . .	101
6.10.4	Learning concentrations and variances . . . . .	102
6.10.5	Softmax bug . . . . .	103
<b>7</b>	<b>Conclusion</b>	<b>105</b>
7.1	Summary of contributions . . . . .	105
7.2	Recurring themes . . . . .	106
7.3	Directions for future work . . . . .	107
<b>Bibliography</b>		<b>109</b>

## *Chapter 1*

---

# Text analysis for the social sciences

---

(Parts of this chapter were originally published as O'Connor et al. (2011).)

## 1.1 Introduction

Corpora of news, books, and social media encode human beliefs and culture. While labor-intensive manual content analysis is a well-established method in the social sciences—going back decades or even centuries (Krippendorff, 2012)—interest in automated text analysis has exploded in recent years, since it is impossible for a person to read all of these rapidly growing archives. Automated methods can help measure and discover patterns and themes described in text corpora: records of opinions, events, and ideas held by participants. Recent reviews, tutorials and workshops on automated text analysis range from political science (Grimmer and Stewart, 2013) to the humanities (Shaw, 2012), with a presence at meetings including *New Directions in Analyzing Text as Data*, *Digital Humanities*, and many computer science conferences such as *ACL*, *EMNLP*, *CIKM*, *ICWSM*, etc.

Researchers have considered a great breadth of questions. Just within academic research, there exist many examples from political science to literature analysis (Figure 1.2). In these works, text analysis is a methodological tool used in service of social science or humanistic questions, where the textual data has been created as the outcome of a socially embedded generation process. This is illustrated in Figure 1.1. A text generation process (*Generator*), constrained by variables of social context (*SocialAttributes*), produces linguistic text data (*Text*). This can be thought of as a stochastic function,

$$\text{Generator}(\text{SocialAttributes}) \mapsto \text{Text}$$

This thesis develops several case studies that are instances of analyzing this process. For example, Chapter 4 studies geographic lexical variation, in which an author's social context is represented as their geographic location. The parameters of the text generation process represent how people in different locations choose to talk about different topics or use different words to talk about them. (Maybe this is due to local topics of interest, or to geographic locality of social communities, which guide people's interests and vocabulary.)

Given observed text data, researchers would like to conduct inference to reverse the text generation process and learn about either social variables or the parameters of the generation process (how social context turns into language). For the geographic example, if we have a dataset of authors with both locations and text, we can learn the parameters of the text generation process,

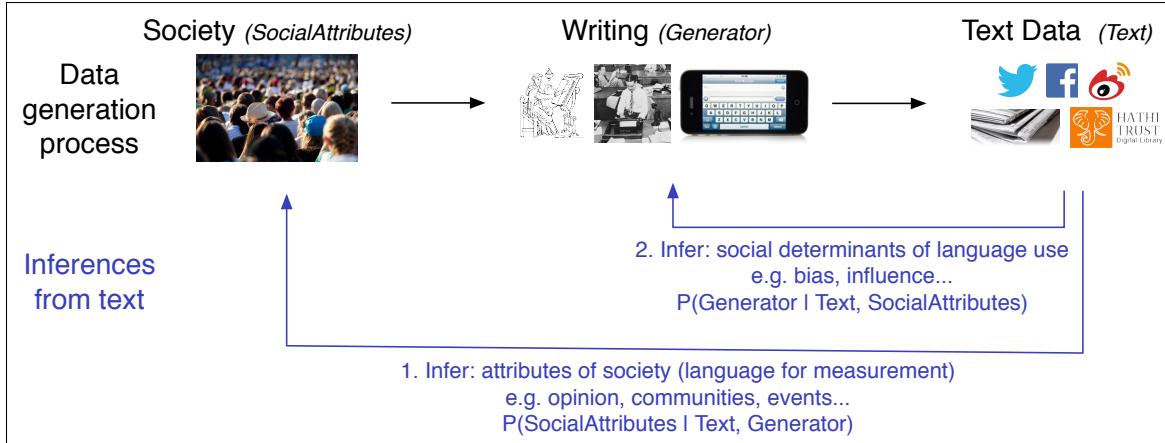


Figure 1.1: Overview: social production of text and the possibilities of text analysis.

which describe the differences in word and topic usage in different locations. We could also compare different hypotheses about this process. Furthermore, given such an inferred model, we can also infer the locations for new authors by analyzing their text.

Using a Bayesian notation for inference among these three groups of variables (*Text*, *Generator*, *SocialAttributes*), there exist two broader classes of text analyses in light of social questions:

### 1. Language for social measurement: $P(\text{SocialAttributes} \mid \text{Text}, \text{Generator})$

The goal here is to infer attributes of society or individuals that are reliably discernible in textual content. For use in such a measurement instrument, text data might have useful properties; for example, it may be cheaper to acquire than surveys or interview methods. This typically requires assumptions about the text generation process. For example, we might take Associated Press articles at their word and extract political events that they are reporting on, and use this as a measurement of how real-world politics changes over time (Chapter 6).

### 2. Language generation as social mechanism: $P(\text{Generator} \mid \text{Text}, \text{SocialAttributes})$

It is critical to understand where the text comes from—the parameters of the text generation process. Language production is deeply embedded in social factors and subject to their biases and influence. Understanding how these factors shape the text is a question in itself. For example, instead of taking news articles at their word, we might analyze how different news sources report things differently—perhaps this is biased by media ownership or political perspectives.

Both types of analysis assume the existence of a statistical relationship between text and social context variables. Text variables may include words, phrases, named entities, or semantic argument structures; social context variables may include the author's community, characteristics, opinions, or a document's embedding in time and space. The parameters of the text generation process encompass social behavior and linguistic production, from individual-level psychological effects (e.g. people's mood might affect how they write), to macro-level trends (e.g. the popularity of topics over time within an academic community) or organizational effects (e.g. economic pressures on publishing companies might affect news coverage).

- Political Science: How do U.S. Senate speeches reflect agendas and attention? How are Senate institutions changing (Quinn et al., 2010)? What are the agendas expressed in Senators' press releases (Grimmer, 2010)? Do U.S. Supreme Court oral arguments predict justices' voting behavior (Black et al., 2011)? Does social media reflect public political opinion, or forecast elections (O'Connor et al., 2010a; Metaxas et al., 2011; Gayo-Avello, 2012)? What determines international conflict and cooperation (Schrodt et al., 1994; King and Lowe, 2003; Shellman, 2008; O'Connor et al., 2013)? How much did racial attitudes affect voting in the 2008 U.S. presidential election (Stephens-Davidowitz, 2012)?
- Economics: How does sentiment in the media affect the stock market (Tetlock, 2007; Lavrenko et al., 2000)? Does sentiment in social media associate with stocks (Gilbert and Karahalios, 2010; Das and Chen, 2007; Bollen et al., 2010)? Do a company's SEC filings predict aspects of stock performance (Kogan et al., 2009; Loughran and McDonald, 2011)? What determines a customer's trust in an online merchant (Archak et al., 2011)? How can we measure macroeconomic variables with search queries and social media text (Askitas and Zimmermann, 2009; Kahn and Kotchen, 2010; O'Connor et al., 2010a)? Can Internet data forecast consumer demand for movies (Asur and Huberman, 2010; Joshi et al., 2010)?
- Psychology: How does a person's mental and affective state manifest in their language (Tausczik and Pennebaker, 2009)? Are diurnal and seasonal mood cycles cross-cultural (Golder and Macy, 2011)?
- Sociology of Science: What are influential topics within a scientific community (Gerrish and Blei, 2010)? What determines a paper's citations (Bethard and Jurafsky, 2010; Ramage et al., 2011; Yogatama et al., 2011)?
- Sociolinguistics: How do geography (Labov et al., 2006; Nerbonne, 2009; Eisenstein et al., 2010), gender (Bamman et al., 2014), class, race, and other social factors (Tagliamonte, 2006; O'Connor et al., 2010b; Eisenstein et al., 2011c) influence linguistic variation, and the lexical diffusion process?
- Public Health: How can search queries and social media help measure levels of the flu and better understand other public health issues (Ginsberg et al., 2009; Culotta, 2010; Paul and Dredze, 2011; Broniatowski et al., 2013)?
- History: How did modern English legal institutions develop over the 17th to 20th centuries (Cohen et al., 2011)? When did concepts of religion, secularism, and social institutions develop over two millennia of Latin literature (Bamman and Crane, 2011)? What do topic labels in a historical encyclopedia reveal about contemporary ways of thought (Horton et al., 2009)?
- Literature: How do demographic determinants of fictional characters affect their language use (Argamon et al., 2009)? Who is the true author of a work of literature or historical documents (Holmes, 1998); for example, Shakespeare (Craig and Kinney, 2009) or the Federalist Papers (Mosteller and Wallace, 1964)?

Figure 1.2: A small sample of social scientific and humanistic questions to which automated text analysis methods have been applied.

While it is useful to distinguish these two types of analysis goals, they are also complementary and can sometimes reuse the same analysis methods. If a model of text can predict social variables (the measurement goal), this gives predictive validity to its text generation parameters (the parameter understanding goal).

## 1.2 Text analysis methods

How should we actually do these analyses, quantitatively speaking? One attractive framework, implied by the notation and description used above, is probabilistic generative modeling: test and develop models of the data-generating process, and use Bayes rule to invert them to infer the quantities of interest. Some of the work in this thesis directly uses this approach for specific aspects of social text generation, like the geographic and demographic topic model in Section 4.2. In the long-term, our research program should strive for this ambitious goal: develop a unified model of human social and linguistic behavior, from which many interesting inferences and predictions can be derived.

At the same time, social science and linguistics are vast and complex fields, and scientific progress must proceed by deductive process of developing and exploring hypotheses concerning many different scientific questions.<sup>1</sup> There exist many fascinating social questions, and a tangle of interesting but often messy text data to query for clues to their answers. Therefore social text analysis should proceed under a *data analytic* approach, incorporating a variety tools from statistics, social science methods, machine learning, and natural language processing. Since these are still early times in the study of language and social behavior, we need to explore a wide variety of methods, and better understand which of them are useful and when. A useful view is to think of automated text analysis as quantitative tools for social science investigations: quantitative methods, alongside traditional qualitative methods, are just another set of tools to investigate core questions of sociology, political science, and economics, and other areas (King et al., 1994).

There are easily dozens of automated text analysis methods that have been used in previous work or could be useful in future social analysis work. In order to help organize this area for practitioners, we suggest thinking about them on three dimensions:

1. Computational and statistical complexity  
(e.g. summary statistics, convex optimization, latent variable learning),
2. Amount of domain assumptions as input  
(e.g. document covariates, hand-built labels, hand-built dictionaries), and
3. Complexity of linguistic representation used in the analysis  
(e.g. word and phrases, entities, opinions, argument structure)

Figure 1.3 schematically illustrates these dimensions. The chapters of this thesis, consisting of tools and case studies, are shown on these dimensions; all of the studies listed at the start of this chapter could be placed on these dimensions as well.

**Domain assumptions** refer to how much knowledge of the substantive issue in question is used in the analysis.

---

<sup>1</sup>While Bayesian notation is useful to describe inferences among broad classes of variables, even if particular statistical techniques are Bayesian, scientific progress may proceed under deductive, not inductive, lines in any case (Gelman and Shalizi, 2013).

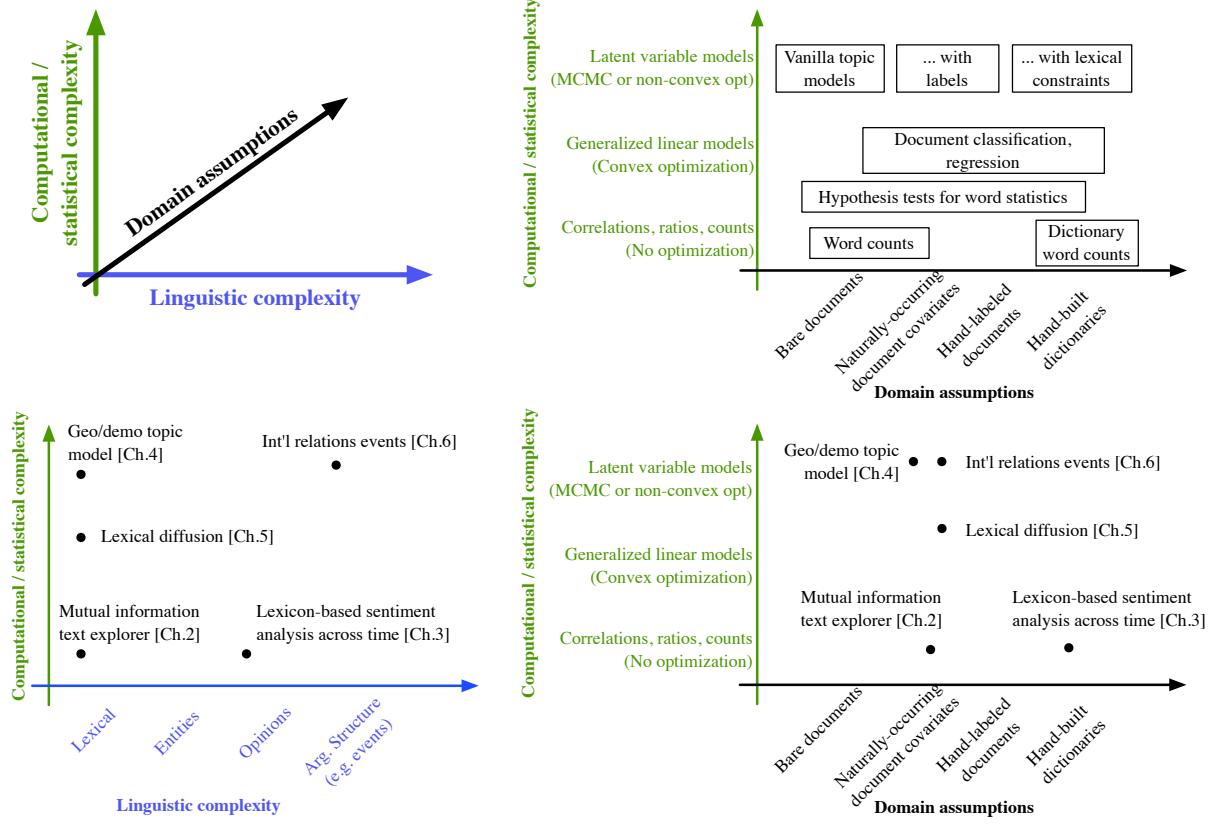


Figure 1.3: **Top:** Taxonomy of analysis methods, with some examples. **Bottom Left:** Work described in this thesis, compared on computation vs. linguistic complexity. **Bottom Right:** Compared on computation vs. domain assumptions.

- A “**bare documents**” analysis only considers the words of documents; for example, examining the most common words in a corpus, or co-occurrence patterns among words.
- **Document covariates** refer to non-textual metadata about the documents, which may have an interesting statistical relationship to the words. Typically this information takes the form of continuous, discrete, or ordinal variables associated with documents or segments of them. The examples in this thesis make extensive use of document covariates that represent social variables, such as time, geography, and demographics. From a strictly computational/statistical perspective, all covariates are not necessarily straightforwardly representative of social processes (for example, in Chapter 2 we use the book of the Bible as a covariate); therefore we use the term *covar* in the rest of this section.
- **Manually labeled documents** may be created in order to better understand particular quantities of interest; for example, annotating news articles with topics or perspectives they describe. Creating and evaluating the codebook (that carefully defines the semantics of the annotations a coder will produce) can be a laborious and iterative process, but is essential to understand the problem and create a reliable coding standard (Krippendorff, 2012). A traditional content analysis project uses manual labeling for all analysis, it can also be combined by computational techniques by relating the labels to textual data. From a strict computational/statistical perspective, a manual label is just another document covariate; the differ-

ence is that these are produced as part of the research process, as opposed to being part of the object under study. (This distinction is not always clear, of course; for example, correcting coding errors may be an essential data cleaning step.)

- Finally, another source of domain information can take the form of **dictionaries**: lists of terms of interest to the analysis, such as names, emotion-indicating words, event-indicating phrases, topic-specific keywords, etc. Ideally, they may be custom-built for the problem (if done manually, a labor-intensive task similar to coding documents), or they may be reused or adapted from already-existing dictionaries (e.g., Freebase (Bollacker et al., 2008) for names or LIWC (Tausczik and Pennebaker, 2009) for affect, though see Grimmer and Stewart (2013)'s critical comments on the naïve use of affect dictionaries). Useful information can be revealed with just a handful of terms; for example, Stephens-Davidowitz (2012) analyzes the Google search query frequencies of one highly charged racial epithet as a proxy for racial attitudes.

The second dimension is the complexity of the **linguistic representation**. Since natural language processing is inherently very difficult, these representations can only be identified imperfectly from text. For more complex representations, NLP systems tend to be less accurate and more computationally expensive. It is useful to catalogue important linguistic phenomena in text, since they comprise the potential objects of analysis.

- **Word and phrase** frequencies, ignoring their order in documents, is the most basic and essential linguistic representation—the “bag of words” representation. Word frequencies capture lexical variation, which is extremely important since words are most basic linguistic units of meaning in text. They are also practical, since words and phrases can often be identified through relatively simple programs for text tokenization.<sup>2</sup> All chapters in this thesis make extensive use of word and phrase frequencies; and they are the primary linguistic representation for Chapters 2–4.
- **Entities** are people, places, or organizations that the text refers to, typically when mentioned by name (a *named entity*). Often, particular actors or entities are the interesting subjects of analysis. Chapter 3, for example, analyzes messages that mention particular political candidates, and Chapter 6 analyzes news articles’ mentions of actors that represent different countries. Depending on the problem, entities might actually *not* be of interest: in Section 5 we seek to analyze the diffusion of novel slang terms, and actually exclude names from the analysis.
- **Opinions** about a topic are held by an opinion holder—perhaps the author of the document (for example, in a product review), or another entity mentioned in the text (Wiebe et al., 2005). One important aspect of an opinion might be a positive or negative opinion. Extracting opinions from text is often called *sentiment analysis*. Opinions are important for many social analysis problems, but are often difficult to reliably extract. Much work in this area uses word frequency approaches, which we apply in Chapter 3, though ongoing work in structured sentiment analysis will be important going forward.
- **Predicate-argument structures** represent the semantic relationships between words in a sentence or document. For example, adjectives modify nouns, representing attributes or aspects

---

<sup>2</sup>Exceptions: some languages, such as Chinese, have a standard orthographic convention that does not use spaces between words, which makes word segmentation a difficult problem. Furthermore, even in a language such as English that usually puts spaces and punctuation between words, in social media and casual text, the orthography is much more complicated, and tokenizers for standard text genres do not work well. We developed and use the tokenization system of O'Connor et al. (2010c); Owoputi et al. (2013) to tackle this. (These works are not part of this thesis.)

of a concept; and verbs have subjects and objects, which sometimes represent entities that initiate and receive the action described by the verb.

A bag-of-words representation fails to describe the interactions between words in a document, whereas argument structures are inherently *relational*, a more complex representation. There are many possible levels of linguistic argument structure to extract from text, such as syntactic dependencies, semantic roles, frame semantics, discourse structure, and events. Chapter 6 uses a syntactic dependencies approach to extract basic event argument structures from text in order to analyze political events described in the news, relating them to social variables of who the actors are and the temporal context of the action.

Finally, the third dimension is **computational and statistical complexity**. Some methods only involve counting words and reporting summary statistics of them; Chapter 2's *MiTextExplorer* tool is interactive software within that paradigm, and Chapter 3's sentiment analysis of Twitter utilizes this level of complexity as well. Other approaches involve more computationally intensive activity such as learning latent semantic representations of words or more complex linguistic units. Chapters 4 and 6 pursue this approach. Many considerations may guide which methods are useful at different stages of an analysis; §4.4 contains some discussion.

In the following chapters, we make use of a wide variety of methods across these dimensions. Easy-to-interpret and computationally cheap methods are often very useful at the start of a project, when iterative exploration of a dataset is essential to gain a basic understanding; but more elaborate models can give further insights, or better target more refined hypotheses. And some questions simply require more computational or linguistic complexity to reach the questions of interest. Further discussion is contained within the chapters themselves and the conclusion (Chapter 7).

### 1.3 Thesis statement

We claim that automatic analysis of text corpora can reveal important attributes of society, through statistical analysis and modeling of linguistic data. Text data can be used to predict or measure social variables, and it can also demonstrate how socially embedded processes guide language production. This is illustrated through several case studies addressing questions in sociolinguistics and political science. While a variety of methods are necessary to accomplish this type of data analysis, we find that latent variable models, probabilistic graphical models, and natural language processing are all crucial tools.

The conclusion chapter contains a summary of contributions in this thesis.

## Chapter 2

# MiTextExplorer: Interactive exploratory text and covariate analysis

(This chapter was originally published as O'Connor (2014).)

### 2.1 Introduction

In this chapter we describe a preliminary experimental system, MiTEXTEXPLORER, for *textual linked brushing*, which allows an analyst to interactively explore statistical relationships between (1) terms, and (2) document metadata (covariates). An analyst can graphically select documents

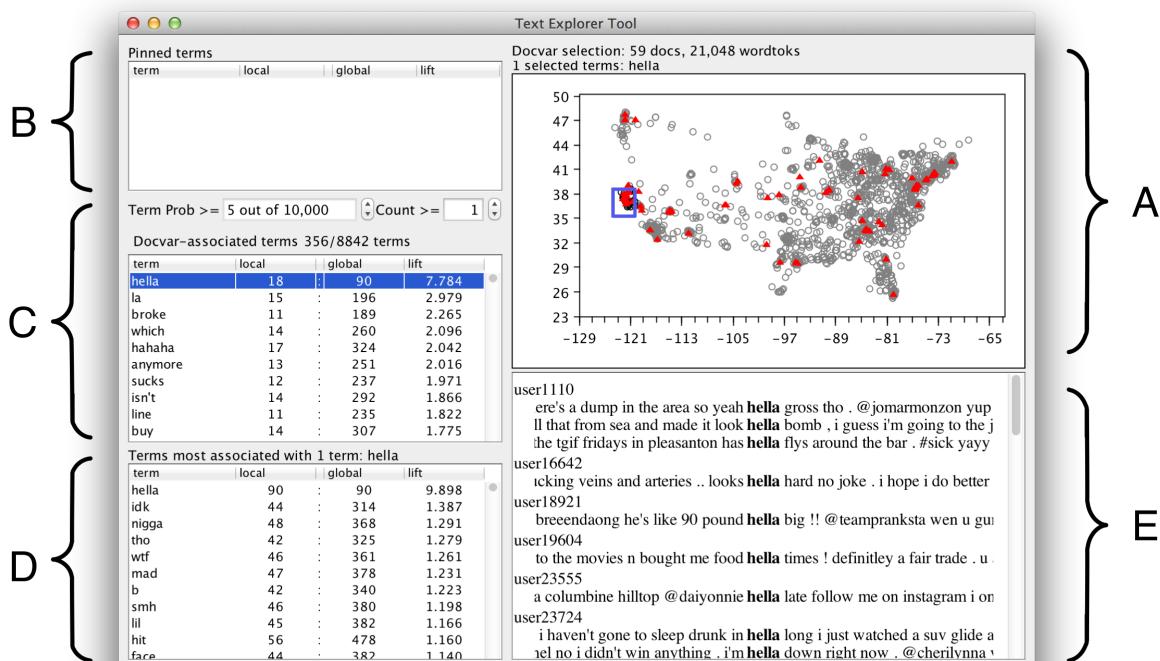


Figure 2.1: Screenshot of MiTEXTEXPLORER, analyzing geolocated tweets.

embedded in a temporal, spatial, or other continuous space, and the tool reports terms with strong statistical associations for the region. The user can then drill down to specific term and term groupings, viewing further associations, and see how terms are used in context. The goal is to rapidly compare language usage across interesting document covariates.

We illustrate examples of using the tool on several datasets, including geo-located Twitter messages, which are more extensively investigated in Chapter 4. We leave a more extensive evaluation for future work.

## 2.2 Motivation: Can we “just look” at statistical text data?

*Exploratory data analysis* (EDA) is an approach to extract meaning from data, which emphasizes learning about a dataset through an iterative process of many analyses which suggest and refine possible hypotheses. It is vital in early stages of a data analysis for data cleaning and sanity checks, which are crucial to help ensure a dataset will be useful. Exploratory techniques can also suggest possible hypotheses or issues for further investigation.

The classical approach to EDA, as pioneered in works such as Tukey (1977) and Cleveland (1993) (and other work from the Bell Labs statistics group during that period) emphasizes visual analysis under nonparametric, model-free assumptions, in which visual attributes are a fairly direct reflection of numerical or categorical aspects of data. As a simple example, consider the well-known Anscombe Quartet (1973), a set of four bivariate example datasets. The Pearson correlation, a very widely used measure of dependence that assumes a linear Gaussian model of the data, finds that each dataset has an identical amount of dependence ( $r = 0.82$ ). However, a scatterplot instantly reveals that very different dependence relationships hold in each dataset (Figure 2.2). The scatterplot is possibly the simplest visual analysis tool for investigating the relationship between two variables, in which the variables’ numerical values are mapped to horizontal and vertical space. While the correlation coefficient is a model-based analysis tool, the scatterplot is model-free (or at least, it is effective under an arguably wider range of data generating assumptions), which is crucial for this example.

This nonparametric, visual approach to EDA has been encoded into many data analysis packages, including the now-ubiquitous R language (R Core Team, 2013), which descends from earlier software by the Bell Labs statistics group (Becker and Chambers, 1984). In R, tools such as histograms, boxplots, barplots, dotplots, mosaicplots, etc. are built-in, basic operators in the language. (Wilkinson (2006)’s grammar of graphics more extensively systematizes this approach; see also Wickham (2010); Bostock et al. (2011).)

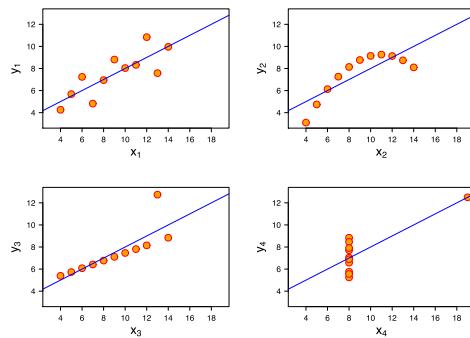


Figure 2.2: Anscombe Quartet. (Source: Wikipedia)

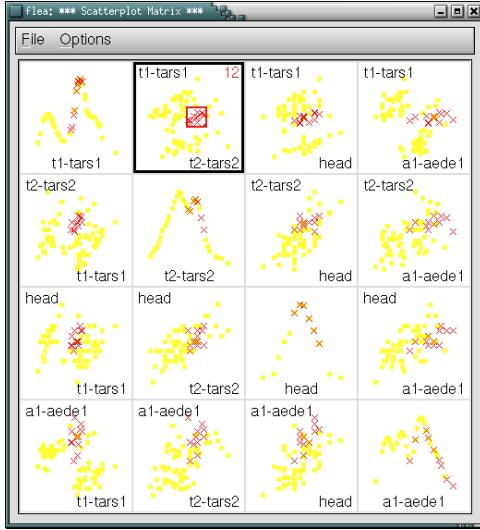


Figure 2.3: Linked brushing with the analysis software *GGobi*. More references at source: [http://www.infovis-wiki.net/index.php?title=Linking\\_and\\_Browsing](http://www.infovis-wiki.net/index.php?title=Linking_and_Browsing)

In the meantime, *textual data* has emerged as a resource of increasing interest for many scientific, business, and government data analysis applications. Consider the use case of automated content analysis (a.k.a. text mining) as a tool for investigating social scientific and humanistic questions (Grimmer and Stewart, 2013; Jockers, 2013; Shaw, 2012; O'Connor et al., 2011). The content of the data is under question: analysts are interested in what/when/how/by-whom different concepts, ideas, or attitudes are expressed in a corpus, and the trends in these factors across time, space, author communities, or other document-level covariates (often called metadata). Comparisons of word statistics across covariates are essential to many interesting questions or social measurement problems, such as

- What topics tend to get censored by the Chinese government online, and why (Bamman et al., 2012; King et al., 2013)? *Covariates*: whether a message is deleted by censors, time/location of message.
- What drives media bias? Do newspapers slant their coverage in response to what readers want (Gentzkow and Shapiro, 2010)? *Covariates*: political preferences of readers, competitiveness of media markets.

In this work, we focus on the question: What should be the baseline exploratory tools for textual data, to discover important statistical associations between *text* and *document covariates*? Ideally, we'd like to "just look" at the data, in the spirit of scatterplotting the Anscombe Quartet. An analysis tool to support this should not require any statistical model assumptions, and should display the data in as direct a form as possible.

For low-dimensional, non-textual data, the base functionality of R prescribes a broad array of useful defaults: one-dimensional continuous data can be histogrammed (`hist(x)`), or kernel density plotted (`plot(density(x))`), while the relationship between two dimensions of continuous variables can be viewed as a scatterplot (`plot(x,y)`); or perhaps a boxplot for discrete *x* and continuous *y* (`boxplot(x,y)`); and so on. Commercial data analysis systems such as Excel, Stata, Tableau, JMP, StatWing, etc., have similar functionality.

These visual tools can be useful for analyzing derived content statistics from text—for example, showing a high-level topic or sentiment frequency trending over time—but they cannot visualize the text itself. Text data consists of a linear sequence of high-dimensional discrete variables (words). The most aggressive and common analysis approach, bag-of-words, eliminates the problematic sequential structure, by reducing a document to a high-dimensional discrete counts over words. But still, none of the above visual tools makes sense for visualizing a word distribution; many popular tools simply crash or become very slow when given word count data. And besides the issues of discrete high-dimensionality, text is unique in that it has to be manually *read* in order to more reliably understand its meaning. Natural language processing tools can sometimes extract partial views of text meaning, but full understanding is a long ways off; and the quality of available NLP tools varies greatly across corpora and languages. A useful exploratory tool should be able to work with a variety of levels of sophistication in NLP tooling, and allow the user to fall back to manual reading when necessary. The tool should support Shneiderman (1996)'s recommendation of *overview first, zoom and filter, then details-on-demand*; the most detailed view of text data, of course, is to support reading of individual snippets and documents

## 2.3 MiTEXTEXPLORER: linked brushing for text and covariate correlations

The analysis tool presented here, MiTEXTEXPLORER, is designed for exploratory analysis of relationships between document covariates—such as time, space, or author community—against textual variables—words, or other units of meaning, that can be counted per document. Unlike topic model approaches to analyzing covariate-text relationships (Mimno, 2012; Roberts et al., 2013), there is no dimension reduction of the terms. Instead, interactivity allows a user to explore more of the high-dimensional space, by specifying a *document selection* ( $Q$ ) and/or a *term selection* ( $T$ ). We are inspired by the *linking and brushing* family of techniques in interactive data visualization, in which an analyst can select a group of data points under a query in one covariate space, and see the same data selection in a different covariate space (Figure 2.3; see Buja et al. (1996), and e.g. Becker and Cleveland (1987); Buja et al. (1991); Martin and Ward (1995); Cook and Swayne (2007)). In our case, one of the variables is text.

The interface consists of several *linked views*, which contain:

- (A) a view of the documents in a two-dimensional covariate space (e.g. scatterplot),
- (B) an optional list of pinned terms,
- (C) *document-associated terms*: a view of the relatively most frequent terms for the current document selection,
- (D) *term-associated terms*: a view of terms that relatively frequently co-occur with the current term selection; and
- (E) a keyword-in-context (KWIC) display of textual passages for the current term selection.

Figure 2.1 shows the interface viewing a corpus of 201,647 geo-located Twitter messages from 2,000 users during 2009-2012, which have been tagged with their author's spatial coordinates through a mobile phone client and posted publicly; for data analysis, their texts have been lowercased and tokenized appropriately (Owoputi et al., 2013; O'Connor et al., 2010c). Since this type of corpus

contains casual, everyday language, it is a dataset that may illuminate geographic patterns of slang and lexical variation in local dialects (Eisenstein et al., 2012, 2010).

The document covariate display (A) uses (longitude, latitude) positions as the 2D space. The corpus has been preprocessed to define a document as the concatenation of messages from a single author, with its position the average location of the author’s messages. When the interface loads, all points in (A) are initially gray, and all other panels are blank.

### 2.3.1 Covariate-driven queries

A core interaction, *brushing*, consists of using the mouse to select a rectangle in the (x,y) covariate space. Figure 2.1 shows a selection around the Bay Area metropolitan area (blue rectangle). Upon selection, the document-driven term display (C) is updated to show the relatively most frequent terms in the document selection. Let  $Q$  denote the set of documents that are selected by the current covariate query. The tool ranks terms  $w$  by their (exponentiated) pointwise mutual information, a.k.a. *lift*, for  $Q$ :

$$\text{lift}(w; Q) = \frac{p(w|Q)}{p(w)} \quad \left( = \frac{p(w, Q)}{p(w)p(Q)} \right) \quad (2.1)$$

This quantity measures how much more frequent the term is in the queryset, compared to the baseline global probability in the corpus ( $p(w)$ ). Probabilities are calculated with simple MLE relative frequencies, i.e.

$$\frac{p(w|Q)}{p(w)} = \frac{\sum_{d \in Q} n_{dw}}{\sum_{d \in Q} n_d} \frac{N}{n_w} \quad (2.2)$$

where  $d$  denotes a document ID,  $n_{dw}$  the count of word  $w$  in document  $d$ , and  $N$  the number of tokens in the corpus. PMI gives results that are much more interesting than results from ranking  $w$  on raw probability within the query set ( $p(w|Q)$ ), since that simply shows grammatical function words or other terms that are common both in the queryset and across the corpus, and not distinctive for the queryset.<sup>1</sup>

A well-known weakness of PMI is over-emphasis on rare terms; terms that appear only in the queryset, even if they appear only once, will attain the highest PMI value. One way to address this is through a smoothing prior/pseudocounts/regularization, or through statistical significance ranking (see §2.4). For simplicity, we use a minimum frequency threshold filter. The user interface allows minimums for either local or global term frequencies, and to easily adjust them, which naturally shifts the emphasis between specific and generic language. All methods to protect against rare probabilistic events necessarily involve such a tradeoff parameter that the user ought to experiment with; given this situation, we might prefer a transparent mechanism instead of mathematical priors (though see also §2.4).

Figure 2.1 shows that *hella* is the highest ranked term for this spatial selection (and frequency threshold), occurring 7.8 times more frequently compared to the overall corpus; this comports with surveyed intuitions of Californian English speakers (Bucholtz et al., 2007). For full transparency to the user, the local and global term counts are shown in the table. (Since *hella* occurred 18 times in the queryset and 90 times globally, this implies the simple conditional probability  $p(Q|w) = 18/90$ ; and indeed, ranking on  $p(Q|w)$  is equivalent to ranking on PMI, since exponentiated PMI is  $p(Q|w)/p(Q)$ .) The user can also sort by local count to see the raw most-frequent

---

<sup>1</sup>The term “lift” is used in business applications (Provost and Fawcett, 2013), while PMI has been used in many NLP applications to measure word associations.

user1110	guess i'm going to the jungle ( <b>la</b> ) @killa_kimbo its totally tru sh" (@seanygrey i will be in <b>la</b> by morning :) that's a fuckin
user29006	iper . @gastelo12 did u bust ? <b>la</b> ! la laa laa la la laa . god . @gastelo12 did u bust ? la ! <b>la</b> laa laa la la laa . goodm @gastelo12 did u bust ? la ! la <b>la</b> laa laa la la laa . goodmorn 2 did u bust ? la ! la laa laa <b>la</b> la laa . goodmorning my li did u bust ? la ! la laa laa <b>la</b> la laa . goodmorning my little 1 u bust ? la ! la laa laa la la <b>la</b> laa . goodmorning my little r
user31473	me @cherylsatijpto ;) balik dr <b>la</b> kpn ? bb is a distraction , it k
user34771	y twiin sister is going to be in <b>la</b> for my bros middleschool gr
user47627	san fracisco is way better than <b>la</b> trust me . :) @teammahone y
user5149	king you a nuisance . i'll be in <b>la</b> this weekend hobnobbing w yself right now . just drove to <b>la</b> from sf and back alone for th
user5239	co @jorge_cortesc en pipolos <b>la</b> comida esta super grasosa #t @s me voy a dormir aca ya es <b>la</b> 1am supongo q alla las 3am

Figure 2.4: KWIC examples of “la” usage in tweets selected in Figure 2.1.

term report for the document selection. As the user reshapes the query box, or drags it around the space, the terms in panel (C) are updated.

Not shown are options to change the term frequency representation. For exposition here, probabilities are formulated as counts of tokens, but this can be problematic for social media data, since a single user might use a term a very large number of times. The above analysis is conducted with an indicator representation of terms per user, so all frequencies refer to the probability that a user uses the term at least once. However, the other examples in this paper use token-level frequencies, which seem to work fine. It is an interesting statistical analysis question how to derive a single range of methods to work across these situations.

### 2.3.2 Term selection and KWIC views

Terms in the table (C) can be clicked and selected, forming a term selection as a set of terms  $T$ . This action drives several additional views:

- (A) documents containing the term are highlighted in the document covariate display (here, in red),
- (E) examples of the term’s usage, in Keyword-in-Context style with vertical alignment for the query term; and
- (D) other terms that frequently co-occur with  $T$  (§2.3.3).

The KWIC report in (E) shows examples of term’s usage. For example, why is the term “la” in the PMI list? My initial thought was that this was an example of “LA”, short for “Los Angeles”. But clicking on “la” instantly disproves this hypothesis—Figure 2.4, showing the Los Angeles sense, but also the “la la la” sense, as well as the Spanish function word.

The KWIC alignment makes it easier to rapidly browse examples, and think about a rough assessment of their word sense or how they are used. Figure 2.5 compares how the term “God”

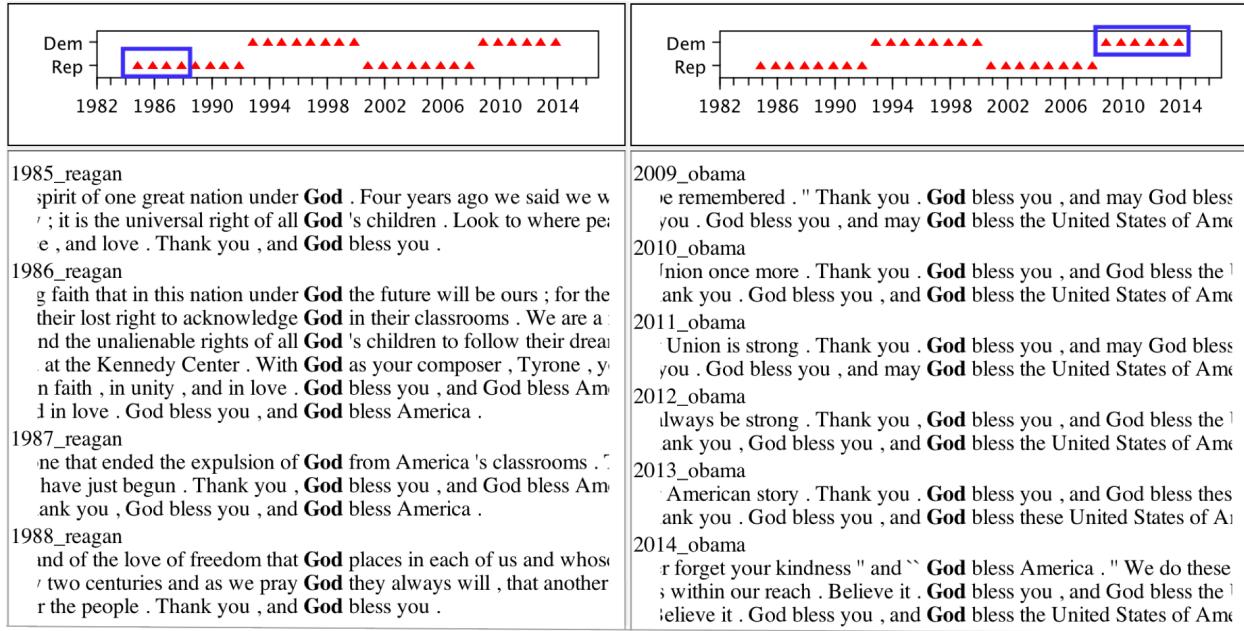


Figure 2.5: KWIC examples of “God” in speeches by Reagan versus Obama.

is used by U.S. presidents Ronald Reagan and Barack Obama, in a corpus of State of the Union speeches, from two different displays of the tool. The predominant usage is the invocation of “God bless America” or similar, nearly ornamental, expressions, but Reagan also has substantive usages, such as references to the role of religion in schools. The vertical alignments of the right-side context words makes it easy to see the “God bless” word sense. We initially found this example simply by browsing the covariate space, and noticing “god” as a frequent term for Reagan, though still occurring for other presidents; the KWIC drilldown better illuminated these distinctions, and suggests differences in political ideologies between the presidents.

In many exploratory text analysis works, especially when utilizing topic models (see Chapter 5), it is common to look at word lists produced by a statistical analysis method and think about what they might mean. At least in our experience doing this, We have often found that seeing examples of words in context has disproved my initial intuitions—for example, in the social media topic models of Chapter 4, it was sometimes difficult to understand the words since we were unfamiliar with the idioms and dialects they came from. Chapter 5 notes the word “ion”, when it appears on Twitter, almost always is a shortened form of “I don’t”, as opposed to the scientific term. The word clusterings induced by a topic model help with this analysis, but reading examples in context is necessary to attain a more complete understanding. Hopefully, supporting this activity in an interactive user interface might make exploratory analysis more effective. Currently, the interface simply shows a sample of in-context usages from the document queryset; it would be interesting to perform grouping and stratified sampling based on local contextual statistics. Summarizing local context by frequencies could be done as a trie visualization (Wattenberg and Viégas, 2008); see §2.5.

### 2.3.3 Term-association queries

When a term is selected, its interaction with covariates is shown by highlighting documents in (B) that contain the term. This can be thought of as another document query: instead of being

specified as a region in the covariate space, is specified as a fragment of the discrete lexical space. As illustrated in much previous work (e.g. Church and Hanks (1990); Turney (2001, 2002)), word-to-word PMI scores can find other terms with similar meanings, or having interesting semantic relationships, to the target term.<sup>2</sup>

This panel ranks terms  $u$  by their association with the query term  $v$ . The simplest method is to analyze the relative frequencies of terms in documents that contain  $v$ ,

$$\text{bool-tt-epmi}(u, v) = \frac{p(w_i = u | v \in \text{supp}(d_i))}{p(w_i = u)}$$

Here, the subscript  $i$  denotes a token position in the entire corpus, for which there is a wordtype  $w_i$  and a document ID  $d_i$ . In this notation, the covariate PMI in 2.3.1 would be  $p(w_i = u | d_i \in Q) / p(w_i = u)$ .  $\text{supp}(d_i)$  denotes the set of terms that occur at least once in document  $d_i$ .

This measure is a very simple extension of the document covariate selection mechanism, and easy to understand. However, it is less satisfying for longer documents, since a larger number of occurrences of  $v$  do not lead to a stronger association score. A possible extension is to consider the joint random event of selecting two tokens  $i$  and  $j$  in the corpus, and ask if the two tokens being in the same document is informative for whether the tokens are the words  $(u, v)$ ; that is, measure  $\text{PMI}[(w_i, w_j) = (u, v); d_i = d_j]$ ,

$$\text{freq-tt-epmi}(u, v) = \frac{p(w_i = u, w_j = v | d_i = d_j)}{p(w_i = u, w_j = v)}$$

In terms of word counts, this expression has the form

$$\text{freq-tt-epmi}(u, v) = \frac{\sum_d n_{du} n_{dv}}{n_u n_v} \frac{N^2}{\sum_d n_d^2}$$

The right-side term is a normalizing constant invariant to  $u$  and  $v$ . The left-side term is interesting: it can be viewed as a similarity measure, where the numerator is the inner product of the inverted term-document vectors  $n_{.,u}$  and  $n_{.,v}$ , and the denominator is the product of their  $\ell_1$  norms. This is a very similar form as cosine similarity, which is another normalized inner product, except its denominator is the product of the vectors'  $\ell_2$  norms.

Term-to-term associations allow navigation of the term space, complementing the views of terms driven by document covariates. This part of the tool is still at a more preliminary stage of development. One important enhancement would be adjustment of the context window size allowed for co-occurrences; the formulations above assume a context window the size of the document. Medium sized context windows might capture more focused topical content, especially in very long discourses such as speeches; and the smallest context windows, of size 1, should be more like collocation detection (though see §2.4; this is arguably better done with significance tests, not PMI).

### 2.3.4 Pinned terms

The term PMI views of (C) and (D) are very dynamic, which can cause interesting terms to disappear when their supporting query is changed. It is often useful to select terms to be constantly viewed when the document covariate queries change.

---

<sup>2</sup>For finding terms with similar semantic meaning, distributional similarity may be more appropriate (Turney and Pantel, 2010); this could be interesting to incorporate into the software.

Any term can be double-clicked to be moved to the the table of *pinned terms* (B). The set of terms here does not change as the covariate query is changed; a user can fix a set of terms and see how their PMI scores change while looking at different parts of the covariate space. One possible use of term pinning is to manually build up clusters of terms—for example, topical or synonymous term sets—whose aggregate statistical behavior (i.e. as a disjunctive query) may be interesting to observe. Manually built sets of keywords are a very useful form of text analysis; in fact, the WordSeer corpus analysis tool has explicit support to help users create them (Shrikumar, 2013).

## 2.4 Statistical term association measures

There exist many measures to measure the statistical strength of an association between a term and a document covariate, or between two terms. A number of methods are based on significance testing, looking for violations of a null hypothesis that term frequencies are independent. For collocation detection, which aims to find meaningful non-compositional lexical items through frequencies of neighboring words, likelihood ratio (Dunning, 1993) and chi-square tests have been used (see review in Manning and Schütze (1999)). For term-covariate associations, chi-square tests were used by Gentzkow and Shapiro (2010) to find politically loaded phrases often used by members of one political party; this same method is often used as a feature selection method for supervised learning (Guyon and Elisseeff, 2003).

The approach we take here is somewhat different, being a point estimate approach, analyzing the estimated difference (and giving poor results when counts are small). Some related work for topic model analysis, looking at statistical associations between words and latent topics (as opposed to between words and observed covariates in this work) includes Chuang et al. (2012), whose term saliency function measures one word’s associations against all topics; a salient term tends to have most of its probability mass in a small set of topics. The measure is a form of mutual information,<sup>3</sup> and may be useful for our purposes here if the user wishes to see a report of distinctive terms for a group of several different observed covariate values at once. Blei and Lafferty (2009) ranks words per topic by a measure inspired by TFIDF, which like PMI downweights words that are generically common across all topics.

Finally, hierarchical priors and regularizers can also be used; for example, by penalizing the log-odds parameterization of term probabilities (Eisenstein et al., 2011b; Taddy, 2013). These methods are better in that they incorporate both protection against small count situations, while paying attention to effect size, as well as allowing overlapping covariates and regression control variables; but unfortunately, they are more computationally intensive, as opposed to the above measures which all work directly from sufficient count statistics. An association measure that fulfilled all these desiderata would be very useful. For term-covariate analysis, Monroe et al. (2008) contains a review of many different methods, from both political science as well as computer science; they also propose a hierarchical prior method, and to rank by statistical significance via the asymptotic standard error of the terms’ odds ratios. Kilgarriff (2001) reviews word comparison

---

<sup>3</sup>This is apparent as follows, using notation from their section 3.1:

$$\text{saliency}(w) = p(w) \sum_T p(T|w) \log[p(T|w)/p(T)] = \sum_T p(w, T) \log[p(w, T)/[p(w)p(T)]]$$

This might be called a “half-pointwise” mutual information: between a specific word  $w$  and the topic random variable  $T$ . Mutual information is  $\sum_w \text{saliency}(w)$ .

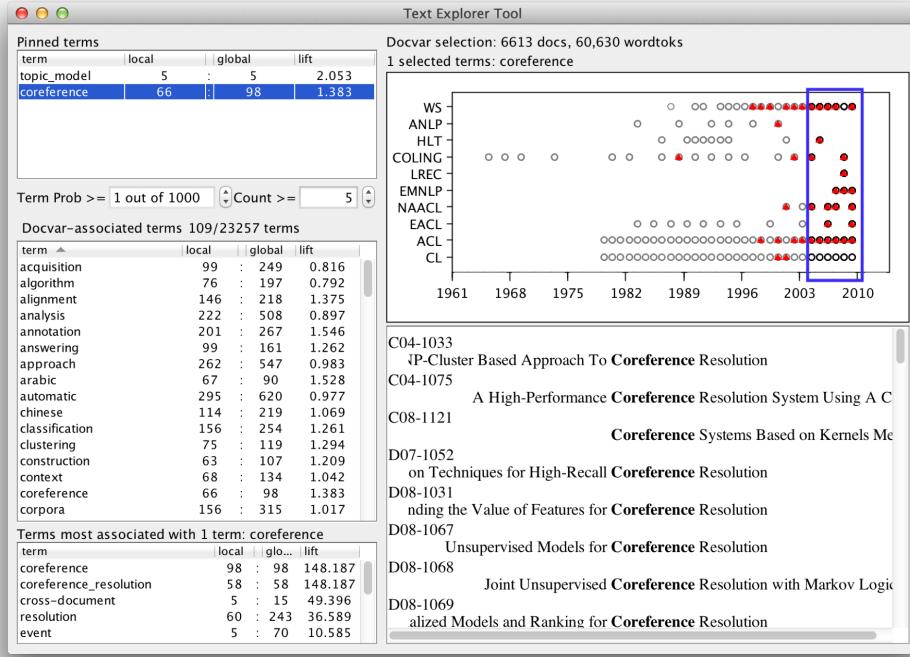


Figure 2.6: MiTEXTEXPLORER for paper titles in the ACL Anthology (Radev et al., 2009). Y-axis is venue (conference or journal name), X-axis is year of publication. Unlike the other figures, docvar-associated terms are sorted alphabetically.

methods as well. The Zeta methods from stylometry (Burrows, 2007; Craig and Kinney, 2009) calculates a word’s distinctiveness based on its variability across a corpus.

Given the large amount of previous work using the significance approach, it merits further exploration for this system.

## 2.5 Related work: Exploratory text analysis

Many systems and techniques have been developed for interactive text analysis. Two such systems, WordSeer and Jigsaw, have been under development for several years, each having had a series of user experiments and feedback. Recent and interesting review papers and theses are available for both of them.

The WordSeer system (Shrikumar, 2013)<sup>4</sup> contains many different interactive text visualization tools, including syntax-based search, and was initially designed for the needs of text analysis in the humanities; the WordSeer 3.0 system includes a word frequency analysis component that can compare word frequencies along document covariates. Interestingly, Shrikumar found in user studies with literary experts that data comparisons and annotation/note-taking support were very important capabilities to add to the system. Unique to the work in this paper is the emphasis on conditioning on document covariates to analyze relative word frequencies, and encouraging the user to change the statistical parameters that govern text correlation measurements. (The

<sup>4</sup><http://wordseer.berkeley.edu/>

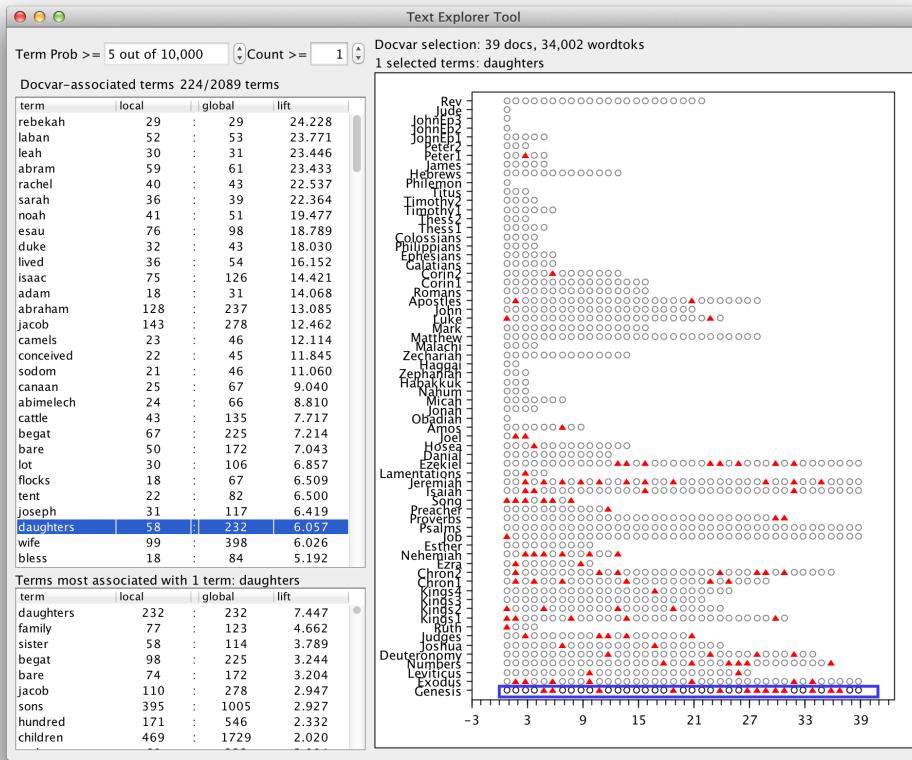


Figure 2.7: MiTEXTEXPLORER for the King James Bible. Y-axis is book, X-axis is chapter (truncated to 39).

term pinning and term-to-term association techniques are certainly less developed than previous work.)

Another text analysis system is Jigsaw (Görg et al., 2013),<sup>5</sup> originally developed for investigative analysis (as in law enforcement or intelligence), which again has many features. It emphasizes visualizations based on entity extractions, such as for names, places, and dates. Görg et al. note that errors in entity extraction were a major problem for users; this might be a worthwhile argument to focus on getting something to first work with simple words/phrases before tackling more complex units of meaning. A section of the review paper is entitled “Reading the documents still matters”, pointing out that analysts did not want just to visualize high-level relationships, but also wanted to read documents in context; this capability was added to later versions of Jigsaw, and supports the emphasis here on the KWIC display.

Both these systems also use variants of Wattenberg and Viégas (2008)’s word tree visualization, which gives a sequential word frequencies as a tree (i.e., what computational linguists might call a trie representation of a high-order Markov model). The “God bless” word sense example from §2.3 indicates that such statistical summarization of local contextual information may be useful to integrate; it is worth thinking how to integrate this against the important need of document covariate analysis, while being efficient with the use of space.

Many other systems, especially ones designed for literary content analysis, emphasize concor-

<sup>5</sup><http://www.cc.gatech.edu/gvu/ii/jigsaw/>

dances and keyword searches within a text; for example, Voyeur/Voyant (Rockwell et al., 2010),<sup>6</sup> which also features some document covariate analysis through temporal trend analyses for individual terms. Another class of approaches emphasizes the use of document clustering or topic models (Gardner et al., 2010; Newman et al., 2010; Grimmer and King, 2011; Chaney and Blei, 2013), while Overview<sup>7</sup> emphasizes hierarchical document clustering paired with manual tagging.

Finally, considerable research has examined exploratory visual interfaces for information retrieval, in which a user specifies an information need in order to find relevant documents or passages from a corpus (Hearst (2009), Ch. 10). Information retrieval problems have some similarities to text-as-data analysis in the need for an exploratory process of iterative refinement, but the text-as-data perspective differs in that it requires an analyst to understand content and contextual factors across multiple or many documents.

## 2.6 Future work

The current MiTEXTEXPLORER system is an extremely simple prototype to explore what sorts of “bare words” text-and-covariates analyses are possible. Several major changes will be necessary for more serious use.

First, essential basic capabilities must be added, such as a search box the user can use to search and filter the term list.

Second, the document covariate display needs to support more than just scatterplots. When there are hundreds or more documents, summarization is necessary in the form of histograms, kernel density plots, or other tools. For example, for a large corpus of documents over time, a lineplot or temporal histogram is more appropriate, where each timestep has a document count. The ACL Anthology scatterplot (Figure 2.6, Radev et al. (2009)), which has hundreds of overplotted points at each (year,venue) position, makes clear the limitations of the current approach.

Better visual feedback for term selections here could be useful—for example, sizing document points monotonically with the term’s frequency (rather than just presence/absence), or using stacked line plots—though certain visual depictions of frequency may be difficult given the Zipfian distribution of word frequencies.

Furthermore, document structures may be thought of as document covariates. A single book has interesting internal variation that could be analyzed itself. Figure 2.7 shows the King James Bible, which has a hierarchical structure of book, chapter, and verse. Here, the (y,x) coordinates represent books and chapters. A more specialized display for book-level structures, or other discourse structures, may be appropriate for book-length texts.

Finally, a major goal of this work is to use analysis methods that can be computed on the fly, but the current prototype only works with small datasets. Hierarchical spatial indexing techniques (e.g. r-trees), may make it possible to interactively compute sums for covariate PMI scoring over very large numbers of documents. Text indexing is also important for term-driven queries and KWIC views. Techniques from ad-hoc data querying systems may be necessary for further scale (e.g. Melnik et al. (2010)).

Many other directions are possible. The prototype tool, as described in §2.3, is available as open-source software at: <http://brenocon.com/mte/>. It is a desktop application written in Java.

---

<sup>6</sup><http://voyant-tools.org/>,  
<http://hermeneuti.ca/voyeur>

<sup>7</sup><https://www.overviewproject.org/> <http://overview.ap.org/>

## *Chapter 3*

---

# Text sentiment and opinion polling

---

(This chapter was originally published as O'Connor et al. (2010a).)

## 3.1 Introduction

If we want to know, say, the extent to which the U.S. population likes or dislikes Barack Obama, a good approach is to ask a random sample of people—that is, take a poll. Survey and polling methodology, extensively developed through the 20th century (Krosnick et al., 2005), gives numerous tools and techniques to accomplish representative public opinion measurement.

With the dramatic rise of text-based social media, millions of people broadcast their thoughts and opinions on a great variety of topics. Can we analyze publicly available data to infer population attitudes in the same manner that public opinion pollsters query a population? If so, then mining public opinion from freely available text content could be a faster and less expensive alternative to traditional polls. Such analysis would also permit us to consider a greater variety of polling questions, limited only by the scope of topics and opinions people broadcast. Extracting the public opinion from social media text provides a challenging and rich context to explore computational models of natural language, motivating new research in computational linguistics.

In this chapter, we connect measures of public opinion derived from polls with sentiment measured from analysis of text from the popular microblogging site Twitter. We explicitly link measurement of textual sentiment in microblog messages through time, comparing to contemporaneous polling data. Surprisingly, summary statistics derived from extremely simple text analysis techniques are demonstrated to correlate with polling data on consumer confidence and political opinion. We find that temporal smoothing is a critically important issue to support a successful model, since the data is highly variable. This suggests there are many future challenges for using social media analysis for opinion tracking.

## 3.2 Data

We begin by discussing the data used in this study: Twitter for the text data, and public opinion surveys from multiple polling organizations.

### 3.2.1 Twitter Corpus

Twitter is a popular microblogging service in which users post messages that are very short: less than 140 characters, averaging 11 words per message. It is convenient for research because there are a very large number of messages, many of which are publicly available, and obtaining them is technically simple compared to scraping blogs from the web.

We use 1 billion Twitter messages posted over the years 2008 and 2009, collected by querying the Twitter API<sup>1</sup> as well as archiving the “Gardenhose” real-time stream. This comprises a roughly uniform sample of public messages, in the range of 100,000 to 7 million messages per day. (The primary source of variation is growth of Twitter itself; its message volume increased by a factor of 50 over this two-year time period.)

Most Twitter users appear to live in the U.S., but we made no systematic attempt to identify user locations or even message language, though our analysis technique should largely ignore non-English messages.

There probably exist many further issues with this text sample; for example, the demographics and communication habits of the Twitter user population probably changed over this time period, which should be adjusted for given our desire to measure attitudes in the general population. There are clear opportunities for better preprocessing and stratified sampling to exploit these data.

### 3.2.2 Public Opinion Polls

We consider several measures of consumer confidence and political opinion, all obtained from telephone surveys to participants selected through random-digit dialing, a standard technique in traditional polling (Chang and Krosnick, 2003).

**Consumer confidence** refers to how optimistic the public feels, collectively, about the health of the economy and their personal finances. It is thought that high consumer confidence leads to more consumer spending, and further relationships with economic activity have been studied (Ludvigson, 2004; Wilcox, 2007). Knowing the public’s consumer confidence is of great utility for economic policy making as well as business planning.

Two well-known surveys that measure U.S. consumer confidence are the Consumer Confidence Index from the Consumer Board, and the Index of Consumer Sentiment (ICS) from the Reuters/University of Michigan Surveys of Consumers.<sup>2</sup> We use the latter, as it is more extensively studied in economics, having been conducted since the 1950s. The ICS is derived from answers to five questions administered monthly in telephone interviews with a nationally representative sample of several hundred people; responses are combined into the index score. Two of the questions, for example, are:

“We are interested in how people are getting along financially these days. Would you say that you (and your family living there) are better off or worse off financially than you were a year ago?”

“Now turning to business conditions in the country as a whole—do you think that during the next twelve months we’ll have good times financially, or bad times, or what?”

We also use another poll, the Gallup Organization’s “Economic Confidence” index,<sup>3</sup> which is derived from answers to two questions that ask interviewees to rate the overall economic health

---

<sup>1</sup>This scraping effort was conducted by Brendan Meeder.

<sup>2</sup>Downloaded from <http://www.sca.isr.umich.edu/>.

<sup>3</sup>Downloaded from <http://www.gallup.com/poll/122840/gallup-daily-economic-indexes.aspx>.

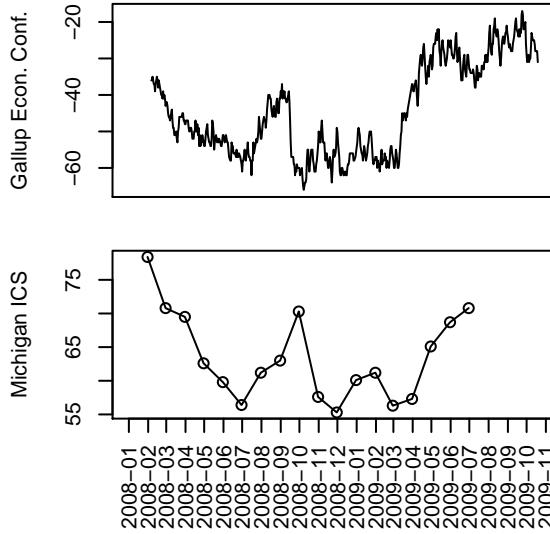


Figure 3.1: Monthly Michigan ICS and daily Gallup consumer confidence poll.

of the country. This only addresses a subset of the issues that are incorporated into the ICS. We are interested in it because, unlike the ICS, it is administered daily (reported as three-day rolling averages). Frequent polling data are more convenient for our comparison purpose, since we have fine-grained, daily Twitter data, but only over a two-year period. Both datasets are shown in Figure 3.1.

For **political opinion**, we use two sets of polls. The first is Gallup's daily tracking poll for the presidential job approval rating for Barack Obama over the course of 2009, which is reported as 3-day rolling averages.<sup>4</sup> These data are shown in Figure 3.2.

The second is a set of tracking polls during the 2008 U.S. presidential election cycle, asking potential voters whether they would vote for Barack Obama or John McCain. Many different organizations administered them throughout 2008; we use a compilation provided by Pollster.com, consisting of 491 data points from 46 different polls.<sup>5</sup> The data are shown in Figure 3.3.

### 3.3 Text Analysis

From text, we are interested in assessing the population's aggregate opinion on a topic. Immediately, the task can be broken down into two subproblems:

1. Message retrieval: identify messages relating to the topic.
2. Opinion estimation: determine whether these messages express positive or negative opinions or news about the topic.

If there is enough training data, this could be formulated as a topic-sentiment model (Mei et al., 2007), in which the topics and sentiment of documents are jointly inferred. Our dataset, however,

<sup>4</sup>Downloaded from <http://www.gallup.com/poll/113980/Gallup-Daily-Obama-Job-Approval.aspx>.

<sup>5</sup>Downloaded from <http://www.pollster.com/polls/us/08-us-pres-ge-mvo.php>

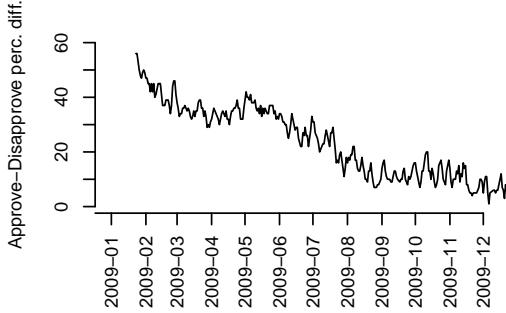


Figure 3.2: 2009 presidential job approval (Barack Obama).

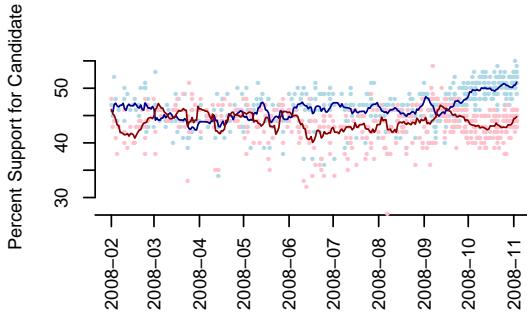


Figure 3.3: 2008 presidential elections, Obama vs. McCain (blue and red). Each poll provides separate Obama and McCain percentages (one blue and one red point); lines are 7-day rolling averages.

is asymmetric, with millions of text messages per day (and millions of distinct vocabulary items) but only a few hundred polling data points in each problem. It is a challenging setting to estimate a useful model over the vocabulary and messages. The signal-to-noise ratio is typical of information retrieval problems: we are only interested in information contained in a small fraction of all messages.

We therefore opt to use a transparent, deterministic approach based on prior linguistic knowledge, counting instances of positive-sentiment and negative-sentiment words in the context of a topic keyword.

### 3.3.1 Message Retrieval

We only use messages containing a topic keyword, manually specified for each poll:

- For consumer confidence, we use *economy*, *job*, and *jobs*.
- For presidential approval, we use *obama*.
- For elections, we use *obama* and *mccain*.

Each topic subset contained around 0.1–0.5% of all messages on a given day, though with occasional spikes, as seen in Figure 3.4. These appear to be driven by news events. All terms have

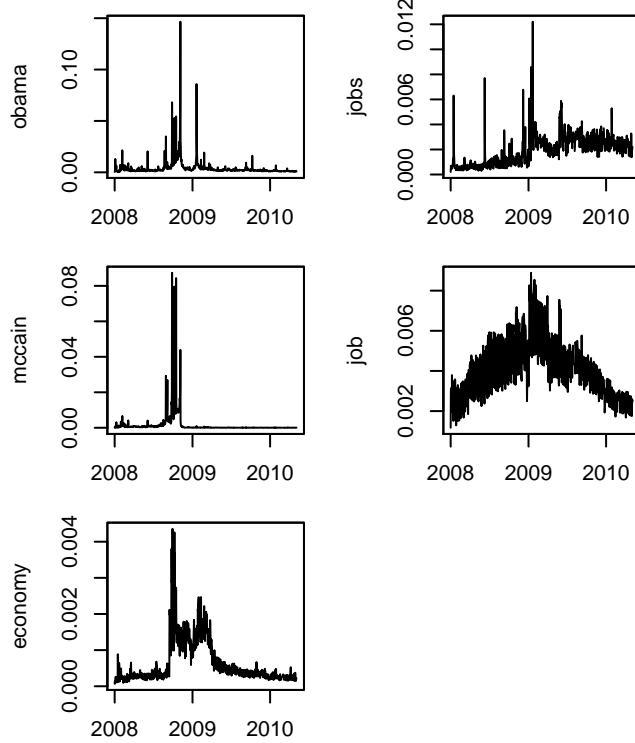


Figure 3.4: Fraction of Twitter messages containing various topic keywords, per day.

a weekly cyclical structure, occurring more frequently on weekdays, especially in the middle of the week, compared to weekends. (In the figure, this is most apparent for the term *job* since it has fewer spikes.) Nonetheless, these fractions are small. In the earliest and smallest part of our dataset, the topic samples sometimes come out just several hundred messages per day; but by late 2008, there are thousands of messages per day for most datasets.

### 3.3.2 Opinion Estimation

We derive day-to-day sentiment scores by counting positive and negative messages. Positive and negative words are defined by the subjectivity lexicon from OpinionFinder, a word list containing about 1,600 and 1,200 words marked as positive and negative, respectively (Wilson et al.).<sup>6</sup> We do not use the lexicon’s distinctions between weak and strong words.

A message is defined as positive if it contains any positive word, and negative if it contains any negative word. (This allows for messages to be both positive and negative.) This gives similar results as simply counting positive and negative words on a given day, since Twitter messages are so short (about 11 words).

We define the sentiment score  $x_t$  on day  $t$  as the ratio of positive versus negative messages on the topic, counting from that day’s messages:

$$\begin{aligned} x_t &= \frac{\text{count}_t(\text{pos. word} \wedge \text{topic word})}{\text{count}_t(\text{neg. word} \wedge \text{topic word})} \\ &= \frac{p(\text{pos. word} \mid \text{topic word}, t)}{p(\text{neg. word} \mid \text{topic word}, t)} \end{aligned} \tag{3.1}$$

<sup>6</sup> Available at <http://www.cs.pitt.edu/mpqa>.

where the likelihoods are estimated as relative frequencies.

We performed casual inspection of the detected messages and found many examples of falsely detected sentiment. For example, the lexicon has the noun *will* as a weak positive word, but since we do not use a part-of-speech tagger, this causes thousands of false positives when it matches the verb sense of *will*.<sup>7</sup> Furthermore, recall is certainly very low, since the lexicon is designed for well-written standard English, but many messages on Twitter are written in an informal social media dialect of English, with different and alternately spelled words, and emoticons as potentially useful signals. Creating a more comprehensive lexicon with distributional similarity techniques could improve the system; Velikovich et al. (2010) find that such a web-derived lexicon substantially improves a lexicon-based sentiment classifier.

### 3.3.3 Comparison to Related Work

The sentiment analysis literature often focuses on analyzing individual documents, or portions thereof (for a review, see Pang and Lee (2008)). Our problem is related to work on sentiment information retrieval, such as the TREC Blog Track competitions that have challenged systems to find and classify blog posts containing opinions on a given topic (Ounis et al., 2008).

The sentiment feature we consider, presence or absence of sentiment words in a message, is one of the most basic ones used in the literature. If we view this system in the traditional light—as subjectivity and polarity detection for individual messages—it makes many errors, like all natural language processing systems. However, we are only interested in *aggregate* sentiment. A high error rate implies the sentiment detector is a noisy measurement instrument. If we have a large number of measurements, and the cause of sentiment errors is not confounded with the substantive comparisons we want to make (see §3.6), these errors will cancel out relative to the quantity of interest: aggregate public opinion as it changes over time. Furthermore, as Hopkins and King (2010) demonstrate, it can actually be inaccurate to naïvely use standard text analysis techniques, which are usually designed to optimize per-document classification accuracy, when the goal is to assess aggregate population proportions.

Many studies have estimated and made use of aggregated text sentiment. In earlier work, the informal study by Lindsay (2008) focused on lexical induction in building a sentiment classifier for a proprietary dataset of Facebook wall posts (a web conversation/microblog medium broadly similar to Twitter), and demonstrated correlations to several polls conducted during part of the 2008 presidential election. Other earlier uses of aggregate text sentiment time series include analyzing stock behavior based on text from blogs (Gilbert and Karahalios, 2010), news articles (Lavrenko et al., 2000; Koppel and Shrimberg, 2004) and investor message boards (Antweiler and Frank, 2004; Das and Chen, 2007). Bollen et al. (2010) tried to predict stocks from Twitter, but see the critique by Anonymous (2012). Dodds and Danforth (2009) use an emotion word counting technique for purely exploratory analysis of several corpora.

## 3.4 Moving Average Aggregate Sentiment

Day-to-day, the sentiment ratio is volatile. Just like in the topic volume plots (Figure 3.4), the sentiment ratio rapidly rises and falls each day. In order to derive a more consistent signal, and following the same methodology used in public opinion polling, we *smooth* the sentiment ratio with one of the simplest possible temporal smoothing techniques, a moving average over a win-

---

<sup>7</sup>We tried manually removing this and several other frequently mismatching words, but it had little effect.

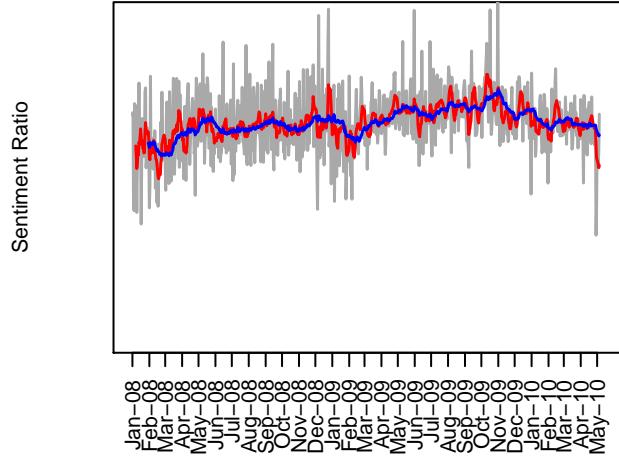


Figure 3.5: Moving average  $MA_t$  of sentiment ratio for *jobs*, under different windows  $k \in \{1, 7, 30\}$ : no smoothing (gray), past week (magenta), and past month (blue). The unsmoothed version spikes as high as 10, omitted for space.

dow of the past  $k$  days:

$$MA_t = \frac{1}{k} (x_{t-k+1} + x_{t-k+2} + \dots + x_t)$$

Smoothing is a critical issue. It causes the sentiment ratio to respond more slowly to recent changes, thus forcing consistent behavior to appear over longer periods of time. Too much smoothing, of course, makes it impossible to see fine-grained changes to aggregate sentiment. See Figure 3.5 for an illustration of different smoothing windows for the *jobs* topic.

### 3.5 Correlation Analysis: Is text sentiment a leading indicator of polls?

Figure 3.6 shows the *jobs* sentiment ratio compared to the two different measures of consumer confidence, Gallup Daily and Michigan ICS. It is apparent that the sentiment ratio captures the broad trends in the survey data. With 15-day smoothing, it is reasonably correlated with Gallup at  $r = 0.731$ . The most glaring difference is a region of high positive sentiment in May-June 2008. But otherwise, the sentiment ratio seems to pick up on the downward slide of consumer confidence through 2008, and the rebound in February/March of 2009.

When consumer confidence changes, can this first be seen in the text sentiment measure, or in polls? If text sentiment responds faster to news events, a sentiment measure may be useful for economic researchers and policymakers. We can test this by looking at leading versions of text sentiment.

First note that the text-poll correlation reported above is the goodness-of-fit metric for fitting slope and bias parameters  $a, b$  in a one variable linear least-squares model:

$$y_t = b + a \left( \frac{1}{k} \sum_{j=0}^{k-1} x_{t-j} \right) + \epsilon_t$$

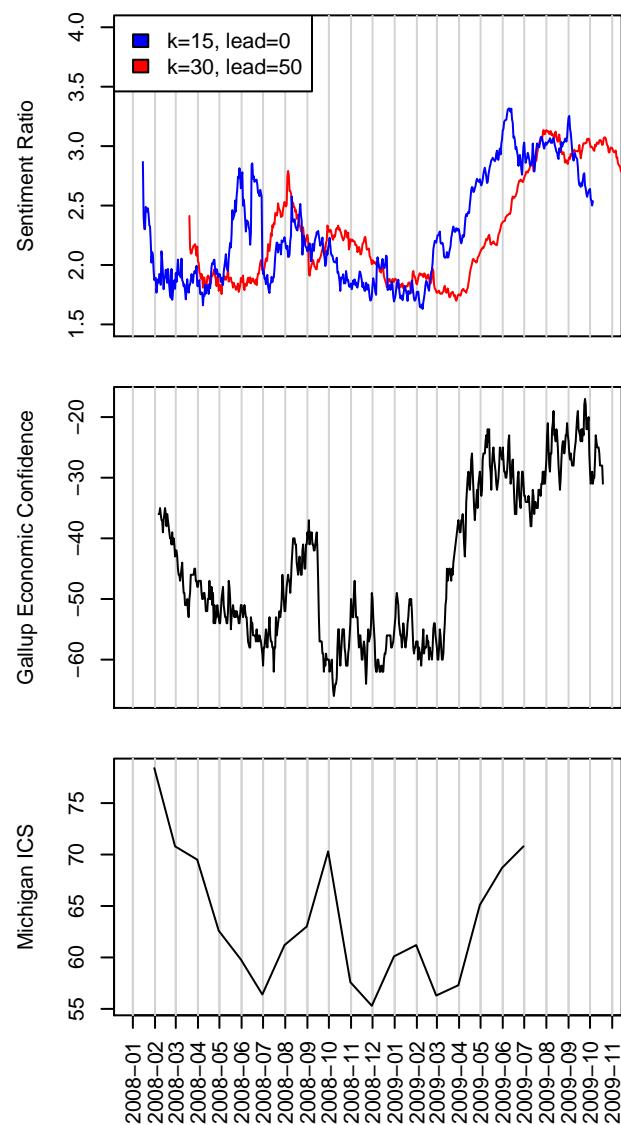


Figure 3.6: Sentiment ratio and consumer confidence surveys. Sentiment information captures broad trends in the survey data.

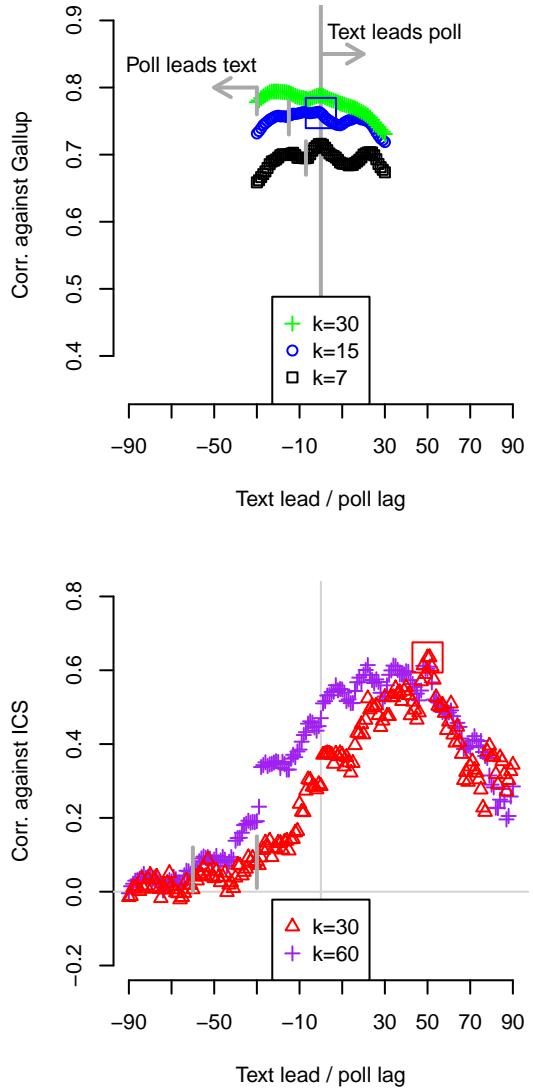


Figure 3.7: Cross-correlation plots: sensitivity to lead and lag for different smoothing windows.  $L > 0$  means the text window completely precedes the poll, and  $L < -k$  means the poll precedes the text. (The window straddles the poll for  $L < -k < 0$ .) The  $L = -k$  positions are marked on each curve. The two parameter settings shown in Figure 3.6 are highlighted with boxes.

for poll outcomes  $y_t$ , daily sentiment ratios  $x_j$ , Gaussian noise  $\epsilon_t$ , and a fixed hyperparameter  $k$ . A poll outcome is compared to the  $k$ -day text sentiment window that ends on the same day as the poll.

The lagged analysis results from introducing a lag hyperparameter  $L$  into the model, so the poll is compared against the text window ending  $L$  days before the poll outcome.

$$y_{t+L} = b + a \left( \frac{1}{k} \sum_{j=0}^{k-1} x_{t-j} \right) + \epsilon_t$$

Graphically, this is equivalent to taking one of the text sentiment lines on Figure 3.6 and shifting it to the right by  $L$  days, then examining the correlation against the consumer confidence polls below.

Polls are typically administered over an interval of time. The ICS is reported once per month (at the end of the month), and Gallup is reported for 3-day windows. We always consider the last day of the poll's window to be the poll date, which is the earliest possible day that the information could be used. Therefore, we would expect both daily measures, Gallup and text sentiment, to always lead ICS, since it measures phenomena occurring over the previous month.

The sensitivity of text-poll correlation to smoothing window and lag parameters  $(k, L)$  is shown in Figure 3.7. The regions corresponding to text preceding or following the poll are marked. Correlation is higher for text leading the poll and not the other way around, so text seems to be a leading indicator. Gallup correlations fall off faster for poll-leads-text than text-leads-poll, and the ICS has similar properties.

If text and polls moved at random relative to each other, these cross-correlation curves would stay close to 0. The fact they have peaks at all strongly suggests that the text sentiment measure captures information related to the polls.

Also note that more smoothing increases the correlation: for Gallup, 7-, 15-, and 30-day windows peak at  $r = 0.716, 0.763$ , and  $0.794$  respectively. The 7-day and 15-day windows have two local peaks for correlation, corresponding to shifts that give alternate alignments of two different humps against the Gallup data, but the better-correlating 30-day window smooths over these entirely. Furthermore, for the ICS, a 60-day window often achieves higher correlation than the 30-day window. These facts imply that the text sentiment information is volatile, and if polls are believed to be a gold standard, then it is best used to detect long-term trends.

It is also interesting to consider ICS a gold standard and compare correlations with Gallup and text sentiment. ICS and Gallup are correlated (best correlation is  $r = 0.864$  if Gallup is given its own smoothing and alignment at  $k = 30, L = 20$ ), which supports the hypothesis that they are measuring similar things, and that Gallup is a leading indicator for ICS. Fixed to 30-day smoothing, the sentiment ratio only achieves  $r = 0.635$  under optimal lead  $L = 50$ . So it is a weaker indicator than Gallup.

Finally, we also experimented with sentiment ratios for the terms *job* and *economy*, which both correlate very poorly with the Gallup poll: 0.10 and 0.07 respectively (with the default  $k = 15, L = 0$ ).<sup>8</sup>

This is a cautionary note on the common practice of stemming words, which in information retrieval can have mixed effects on performance (Manning et al., 2008, Chapter 2). Here, stemming would have conflated *job* and *jobs*, greatly degrading results (to  $r = 0.40$ ).

---

<sup>8</sup>We inspected some of the matching messages to try to understand this result, but since the sentiment detector is very noisy at the message level, it was difficult to understand what was happening.

### 3.5.1 Obama 2009 Job Approval and 2008 Elections

We analyze the sentiment ratio for *obama* and compared it to two series of polls, presidential job approval in 2009, and presidential election polls in 2008, as seen in Figure 3.8. The job approval

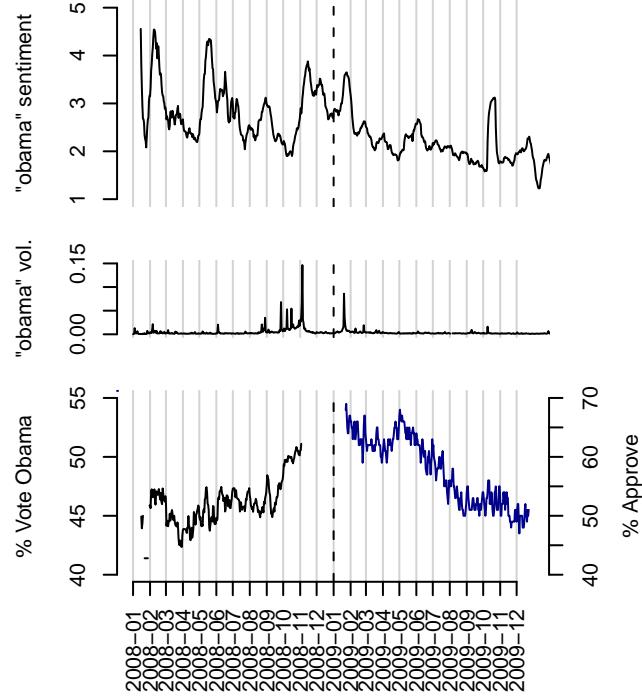


Figure 3.8: The sentiment ratio for *obama* (15-day window), and fraction of all Twitter messages containing *obama* (day-by-day, no smoothing), compared to election polls (2008) and job approval polls (2009).

poll is the most straightforward, being a steady decline since the start of the Obama presidency, perhaps with some stabilization in September or so. The sentiment ratio also generally declines during this period, with  $r = 0.725$  for  $k = 15$ .

However, in 2008 the sentiment ratio does not substantially correlate to the election polls ( $r = -0.08$ ); we compare to the percent of support for Obama, averaged over a 7-day window of tracking polls: the same information displayed in Figure 3.3). Lindsay (2008) found that his daily sentiment score was a leading indicator to one particular tracking poll (Rasmussen) over a 100-day period from June–October 2008. Our measure also roughly correlates to the same data ( $r = 0.44$  versus Lindsay's  $r = 0.57$ ), and only at different lag parameters.

The elections setting may be structurally more complex than presidential job approval. In many of the tracking polls, people can choose to answer any *Obama*, *McCain*, *undecided*, *not planning to vote*, and third-party candidates. Furthermore, the name of every candidate has its own sentiment ratio scores in the data. We might expect the sentiment for *mccain* to be vary inversely with *obama*, but they in fact slightly correlate. It is also unclear how they should interact as part of a model of voter preferences.

Another question is to what how topic frequencies relate to polls. This is a complex question. First note that the message volume for *obama*, shown in Figure 3.8, has the usual daily spikes like other words on Twitter shown in Figure 3.4. Some of these spikes are very dramatic; for example, on November 5th, nearly 15% of all Twitter messages in our sample mentioned the word *obama*.

Furthermore, the *obama* message volume substantially correlates to the poll numbers. Even the raw volume has a 0.52 correlation to the polls, and the 15-day window version is up to  $r = 0.79$  (basically stronger than the sentiment ratio's correlation!). This naïvely seems to indicate that media attention is associated with popularity. But the converse is not true for *mccain*; this word's 15-day message volume *also* correlates to higher Obama ratings in the polls ( $r = 0.74$ ). This contradicts an “any press is good press” idea that media attention toward a candidate causes more popularity. There are many possible explanations; relative frequency, for example, may be a better measurement to investigate.

## 3.6 Measurement bias and inference goals

We do not focus on name frequency analysis, but other work has investigated this more thoroughly in other contexts. Tumasjan et al. (2010) analyzed relative name frequency for the 2009 German parliamentary elections, in Twitter messages over several weeks before the election. They found that the proportions of name frequencies of the six major political parties at the time were very similar to the parties' vote shares, reporting a mean absolute difference of less than 2%. Jungherr et al. (2012) replicated and critiqued this work, noting some sensitivity to details of data collection and name selection (among the six parties their replication found a lower correlation; see also Tumasjan et al. (2012)). Furthermore, they also found a remarkable sensitivity to party selection: the upstart Pirate Party, which was not included in the original study, actually had a far higher name frequency than any of the six major parties, but received a smaller vote share than any of them. We show their results in Figure 3.9. This may not be surprising since the Pirate Party was founded on a platform of online civil liberties issues and had an especially active Internet presence; thus its name frequency was not comparable to the mainstream political parties, at least as an indicator of political preferences of the entire voting population.

This episode highlights a major issue of opinion analysis in social media: are the users representative of the greater population in question? A possible explanation for this result is that Twitter was overrepresented with Pirate-supporting users, compared to the greater population. A related issue is possible differences in platform-specific outreach and communication by political parties; in 2009, it may have been the case that mainstream parties were less effective at online engagement, and thus communications concerning them were relatively less frequent.

Generally speaking, the issue of measurement bias depends on the inference goal. If the goal is to assess the relative strengths of political parties, a relative name frequency analysis requires that the causes of party name mentions be similar across parties. This apparently was not the case for the Pirate Party, as might be expected since it was an outlier in many ways: it was founded just a few years before the election; it was not among the major parties commonly assessed by German polling firms; and it had no representation in the parliament at the time. On the other hand, among the six mainstream parties, the correlation to name frequency is quite remarkable; perhaps the process of their communication strategies and media coverage were comparable enough such that there was some mechanism to produce similar frequencies as voters' preferences—maybe users talked about the ones they liked, or maybe media coverage focused on parties in proportion to their support. These and other hypotheses are important to investigate in future work.

Tumasjan et al., 2010, Jungherr et al., 2012, and many other subsequent works such as Metaxas et al. (2011); Gayo-Avello (2012); Huberty (2013) focus on comparisons of popularity between parties, as a way to predict elections. The work in this chapter instead focuses on temporal correlations to tracking polls over time, where the key comparisons are popularity levels between timesteps. This suggests a number of possible reasons for non-correlations. If properties of the

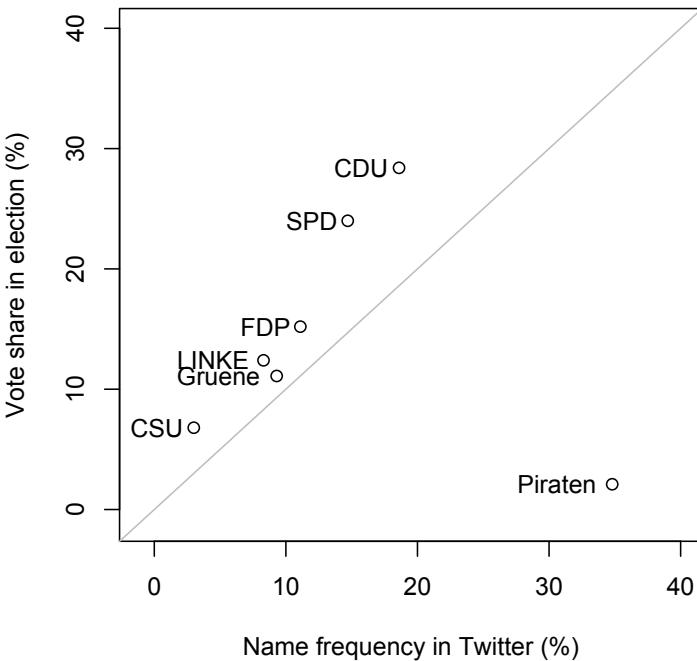


Figure 3.9: Jungherr et al. (2012)'s re-analysis of Tumasjan et al. (2010). *Gruene*=Green Party; *Piraten*=Pirate Party.

user population, such as demographic skew, or how different subpopulations use Twitter (for example, how heavily they use it), changed over time, that would render comparisons across time less meaningful. Even without such change, it could still be problematic since the proportions may not match the overall population; a small shift in opinion in an overrepresented subpopulation will cause an overly large shift to the averaged text measurement.

Besides population representativeness, the process by which the user population generates messages is also fraught: do certain populations write more or less under certain circumstances? Are the messages being observed sent by accounts controlled by personal users, or are they news media or other broadcast-like outlets? Are messages copies or retweets of those; if so, do they still indicate the sender's opinion? (And what about the gray area cases between personal use and broadcasting, such as "microcelebrities" like modestly popular bloggers, or previously unknown people whose messages go viral?) And since social media is monitored as a sign of importance or popularity, interested parties often attempt to manipulate Internet systems to gain apparent popularity (Metaxas and Mustafaraj, 2012).

We have already mentioned issues in the accuracy of the natural language processing techniques. Some of these issues may be invariant to the temporal or other substantively interesting comparisons: for example, if sentiment analyzer fails when negations are used, hopefully the probability of linguistic negation use does not change between timesteps. But linguistic issues may interact with user representativeness: if certain subpopulations or media sources use longer sentences, for example, NLP accuracy may be differentially affected. Chapters 4 and 5 show that

certain demographics, such as minorities and youth, often use non-standard English online, and traditional NLP methods do not adequately capture them.

### 3.7 Conclusion

In the paper we find that a relatively simple sentiment detector based on Twitter data shares some common temporal signal with consumer confidence and presidential job approval polls. While the results do not come without caution, it is encouraging that expensive and time-intensive polling might be supplemented with the simple-to-gather text data that is generated from online social networking. The results suggest that more advanced NLP techniques to improve opinion estimation may be useful.

In this work, we treat polls as a gold standard. Of course, they are also noisy indicators of the truth, subject to biases or variability from question wording, low response rates, cell phone versus landline reachability (AAPOR, 2010), etc. Eventually, future work should seek to understand how these different signals reflect public opinion either as a hidden variable, or as measured from more reliable sources like face-to-face interviews—though currently the sources of measurement error are much better understood for polls than for social media sentiment analysis.

There exist many challenges in trying to derive opinion estimates similar to telephone polling of a more general population. At the very least, poststratification and weighting of the user population should be used: treating opinion analysis of social media data as conditional on user demographics, geography, or other subpopulations which can be reweighted to form an overall picture of the greater population. Wang et al. (2014) use a weighting model of non-representative online polls (from the Microsoft Xbox gaming service) to accurately forecast 2012 U.S. elections. More work is necessary to additionally address issues of variability in communication and linguistic behaviors mentioned in the previous section. A potentially useful technique, when there is enough data, is to use polls as a signal to learn better linguistic representations for classifiers; Beauchamp (2013) finds terms on Twitter that correlate to trends across many U.S. state polls.

Eventually, we see this research progressing to align with the more general goal of query-driven sentiment analysis where one can ask more varied questions of what people are thinking based on text they are already writing. Looking for correspondences with traditional survey data is a potential useful application of sentiment analysis. But it is also a stepping stone toward more complex applications.

## *Chapter 4*

---

# Mixed-membership analysis of geographic and demographic lexical variation in social media

---

## 4.1 Introduction

Even within a single language community, speakers from different backgrounds demonstrate substantial linguistic variation. Salient speaker characteristics include geography (Labov et al., 2006; Kurath, 1949), race (Rickford, 1999), and socioeconomic status (Labov, 1966; Eckert, 1989); they impact language at the phonological, lexical, and morphosyntactic levels (Wolfram and Schilling-Estes, 2005).

Sociolinguistics and dialectology feature a strong quantitative tradition of studying the relationship between language and social and geographical identity (e.g., Labov (1980); Tagliamonte (2006)). In general, these approaches begin by identifying both the communities of interest and the relevant linguistic dimensions of variability; for example, a researcher might identify the term “yinz” as characteristic of Pittsburgh dialect (Dressman, 1979), and then statistically model its relationship to the socioeconomic status of the speaker. This approach requires extensive fieldwork and linguistic expertise to identify the inputs that are to be analyzed. This is particularly challenging in the case of lexical variation, as a large amount of data must be analyzed in order to find demographic patterns for rare lexical items. In terms of Chapter 1’s taxonomy of textual social analysis methods, this approach requires a high amount of domain assumptions—a preselected set of terms or other linguistic markers of dialect.

Given the massive quantities of everyday language that are now available in social media data, a new approach is possible: automatically discover terms that statistically associate with authors’ social variables. In this chapter we develop a model-based, latent variable approach for exploratory analysis of geotagged social media, relating language to geography and demographics. Specifically, we use topic models, which are mixed-membership statistical models of text (Blei et al., 2003; Erosheva et al., 2004). Our model of Twitter users’ language hypothesizes a number of latent topics, each of which is a soft cluster of words, and that each user tends to talk about a subset of these topics. The goal is to learn which words tend to belong to which topics, as well as which topics that individual users tend to use; the model learns both of these latent variables to best explain the data.

The specific model we develop here assumes another latent variable for a user—a community

membership variable—that explains the users’ geographic position (or demographic makeup of their neighborhood), which are tied to community-specific variants of topics. The model thus learns groups of users and terms that are jointly coherent across both linguistic and social data. Besides illustrating the model for exploratory analysis, we additionally validate the geographic model by applying Bayesian inference on its spatio-linguistic representations to predict a user’s geographic location based on the text of their messages. We were surprised by some of the findings in this chapter—in particular, the huge variety of creative neologisms, sometimes very specific to certain geographies or demographics, which are rapidly evolving—which led to, among other work, Chapter 5 on the temporal diffusion of terms.

§4.2 explains our data and model as applied to geography, §4.3 presents an extension to analyze demographics, and §4.4 concludes with reflections and lessons for textual social data analysis.

## 4.2 Geographic topic model and lexical variation

(This section was originally published as Eisenstein et al. (2010).)

### 4.2.1 Synopsis

The rapid growth of geotagged social media raises new computational possibilities for investigating geographic linguistic variation. In this section, we present a multi-level generative model that reasons jointly about latent topics and geographical regions. High-level topics such as “sports” or “entertainment” are rendered differently in each geographic region, revealing topic-specific regional distinctions. Applied to a new dataset of geotagged microblogs, our model recovers coherent topics and their regional variants, while identifying geographic areas of linguistic consistency. The model also enables prediction of an author’s geographic location from raw text, outperforming both text regression and supervised topic models.

#### Introduction

Sociolinguistics and dialectology study how language varies across social and regional contexts. Quantitative research in these fields generally proceeds by counting the frequency of a handful of previously-identified linguistic *variables*: pairs of phonological, lexical, or morphosyntactic features that are semantically equivalent, but whose frequency depends on social, geographical, or other factors (Paolillo, 2002; Chambers, 2009). It is left to the experimenter to determine which variables will be considered, and there is no obvious procedure for drawing inferences from the distribution of multiple variables. In this paper, we present a method for identifying geographically-aligned lexical variation directly from raw text. Our approach takes the form of a probabilistic graphical model capable of identifying both geographically-salient terms and coherent linguistic communities.

One challenge in the study of lexical variation is that term frequencies are influenced by a variety of factors, such as the topic of discourse. We address this issue by adding latent variables that allow us to model topical variation explicitly. We hypothesize that geography and topic interact, as “pure” topical lexical distributions are corrupted by geographical factors; for example, a sports-related topic will be rendered differently in New York and California. Each author is imbued with a latent “region” indicator, which both selects the regional variant of each topic, and generates the author’s observed geographical location. The regional corruption of topics is modeled through a cascade of logistic normal priors—a general modeling approach which we call *cascading* topic models. The resulting system has multiple capabilities, including: (i) analyzing lexical variation

by both topic and geography; (ii) segmenting geographical space into coherent linguistic communities; (iii) predicting author location based on text alone.

This research is only possible due to the rapid growth of social media. Our dataset is derived from the microblogging website Twitter,<sup>1</sup> which permits users to post short messages to the public. Many users of Twitter also supply exact geographical coordinates from GPS-enabled devices (e.g., mobile phones),<sup>2</sup> yielding *geotagged* text data. Text in computer-mediated communication is often more vernacular (Tagliamonte and Denis, 2008), and as such it is more likely to reveal the influence of geographic factors than text written in a more formal genre, such as news text (Labov, 1966).

We evaluate our approach both qualitatively and quantitatively. We investigate the topics and regions that the model obtains, showing both common-sense results (place names and sports teams are grouped appropriately), as well as less-obvious insights about slang. Quantitatively, we apply our model to predict the location of unlabeled authors, using text alone. On this task, our model outperforms several alternatives, including both discriminative text regression and related latent-variable approaches.

#### 4.2.2 Data

The main dataset in this research is gathered from the microblog website Twitter, via its official API. We use an archive of messages collected over the first week of March 2010 from the “Gardenhose” sample stream,<sup>3</sup> which then consisted of 15% of all public messages, totaling millions per day. This short timeframe was chosen in order to obtain a conveniently sized sample that facilitated ease of experimentation. We aggressively filter this stream, using only messages that are tagged with physical (latitude, longitude) coordinate pairs from a mobile client, and whose authors wrote at least 20 messages over this period. We also filter to include only authors who follow fewer than 1,000 other people, and have fewer than 1,000 followers. Kwak et al. (2010) find dramatic shifts in behavior among users with social graph connectivity outside of that range; such users may be marketers, celebrities with professional publicists, news media sources, etc. We also remove messages containing URLs to eliminate bots posting information such as advertising or weather conditions. For interpretability, we restrict our attention to authors inside a bounding box around the contiguous U.S. states, yielding a final sample of about 9,500 users and 380,000 messages, totaling 4.7 million word tokens. We have made this dataset available online.<sup>4</sup>

Informal text from mobile phones is challenging to tokenize; we adapt a publicly available tokenizer<sup>5</sup> originally developed for Twitter (O’Connor et al., 2010c), which preserves emoticons and blocks of punctuation and other symbols as tokens. For each user’s Twitter feed, we combine all messages into a single “document.” We remove word types that appear in fewer than 40 feeds, yielding a vocabulary of 5,216 words. Of these, 1,332 do not appear in the English, French, or Spanish dictionaries of the spell-checking program aspell.

Every message is tagged with a location, but most messages from a single individual tend to come from nearby locations (as they go about their day); for modeling purposes we use only a single geographic location for each author, simply taking the location of the first message in the sample.

The authors in our dataset are fairly heavy Twitter users, posting an average of 40 messages per day (although we see, on average, only 15% of this total). We have little information about

---

<sup>1</sup><http://www.twitter.com>

<sup>2</sup>User profiles also contain self-reported location names, but we do not use that information in this work.

<sup>3</sup>[http://dev.twitter.com/pages/streaming\\_api](http://dev.twitter.com/pages/streaming_api)

<sup>4</sup><http://www.ark.cs.cmu.edu/GeoText>

<sup>5</sup><https://github.com/brendano/tweetmotif>

their demographics, though from the text it seems likely that this user set skews towards teens and young adults. The dataset covers each of the 48 contiguous United States and the District of Columbia.

### 4.2.3 Model

We develop a model that incorporates two sources of lexical variation: topic and geographical region. We treat the text and geographic locations as outputs from a generative process that incorporates both topics and regions as latent variables.<sup>6</sup> During inference, we seek to recover the topics and regions that best explain the observed data.

At the base level of model are “pure” topics (such as “sports”, “weather”, or “slang”); these topics are rendered differently in each region. We call this general modeling approach a *cascading* topic model; we describe it first in general terms before moving to the specific application to geographical variation.

#### Cascading Topic Models

Cascading topic models generate text from a chain of random variables. Each element in the chain defines a distribution over words, and acts as the mean of the distribution over the subsequent element in the chain. Thus, each element in the chain can be thought of as introducing some additional corruption. All words are drawn from the final distribution in the chain.

At the beginning of the chain are the priors, followed by unadulterated base topics, which may then be corrupted by other factors (such as geography or time). For example, consider a base “food” topic that emphasizes words like *dinner* and *delicious*; the corrupted “food-California” topic would place weight on these words, but might place extra emphasis on other words like *sprouts*.

The path through the cascade is determined by a set of indexing variables, which may be hidden or observed. As in standard latent Dirichlet allocation (Blei et al., 2003), the base topics are selected by a per-token hidden variable  $z$ . In the geographical topic model, the next level corresponds to regions, which are selected by a per-author latent variable  $r$ .

Formally, we draw each level of the cascade from a normal distribution centered on the previous level; the final multinomial distribution over words is obtained by exponentiating and normalizing. To ensure tractable inference, we assume that all covariance matrices are uniform diagonal, i.e.,  $a\mathbf{I}$  with  $a > 0$ ; this means we do not model interactions between words.

(§4.2.8 discusses this modeling approach in comparison to related work.)

#### The Geographic Topic Model

The application of cascading topic models to geographical variation is straightforward. Each document corresponds to the entire Twitter feed of a given author during the time period covered by our corpus. For each author, the latent variable  $r$  corresponds to the geographical region of the author, which is not observed. As described above,  $r$  selects a corrupted version of each topic: the  $k$ th basic topic has mean  $\mu_k$ , with uniform diagonal covariance  $\sigma_k^2$ ; for region  $j$ , we can draw the regionally-corrupted topic from the normal distribution,  $\eta_{jk} \sim \mathcal{N}(\mu_k, \sigma_k^2 \mathbf{I})$ .

Because  $\eta$  is normally-distributed, it lies not in the simplex but in  $\mathbb{R}^W$ . We deterministically compute multinomial parameters  $\beta$  by exponentiating and normalizing:  $\beta_{jk} = \exp(\eta_{jk}) / \sum_i \exp(\eta_{jk}^{(i)})$ .

---

<sup>6</sup>The region could be observed by using a predefined geographical decomposition, e.g., political boundaries. However, such regions may not correspond well to linguistic variation.

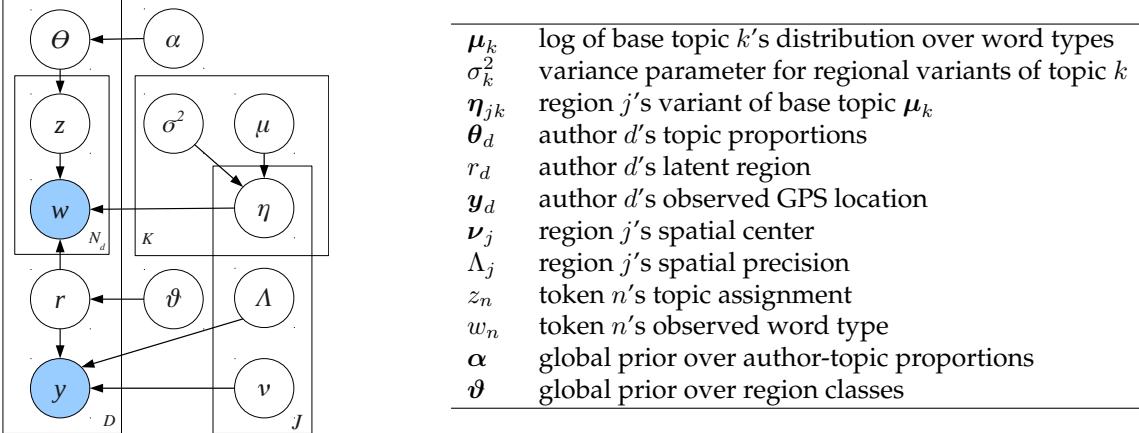


Figure 4.1: Plate diagram for the geographic topic model, with a table of all random variables. Priors (besides  $\alpha$ ) are omitted for clarity, and the document indices on  $z$  and  $w$  are implicit.

This normalization could introduce identifiability problems, as there are multiple settings for  $\eta$  that maximize  $P(w|\eta)$  (Blei and Lafferty, 2006a). However, this difficulty is obviated by the priors: given  $\mu$  and  $\sigma^2$ , there is only a single  $\eta$  that maximizes  $P(w|\eta)P(\eta|\mu, \sigma^2)$ ; similarly, only a single  $\mu$  maximizes  $P(\eta|\mu)P(\mu|a, b^2)$ .

The observed latitude and longitude, denoted  $y$ , are normally distributed and conditioned on the region, with mean  $\nu_r$  and precision matrix  $\Lambda_r$  indexed by the region  $r$ . The region index  $r$  is itself drawn from a single shared multinomial  $\vartheta$ . The model is shown as a plate diagram in Figure 4.1.

Given a vocabulary size  $W$ , the generative story is as follows:

- **Generate base topics:** for each topic  $k < K$ 
  - Draw the base topic from a normal distribution with uniform diagonal covariance:  $\mu_k \sim \mathcal{N}(\mathbf{a}, b^2\mathbf{I})$ ,
  - Draw the regional variance from a Gamma distribution:  $\sigma_k^2 \sim \mathcal{G}(c, d)$ .
- **Generate regional variants:** for each region  $j < J$ ,
  - \* Draw the region-topic  $\eta_{jk}$  from a normal distribution with uniform diagonal covariance:  $\eta_{jk} \sim \mathcal{N}(\mu_k, \sigma_k^2\mathbf{I})$ .
  - \* Convert  $\eta_{jk}$  into a multinomial distribution over words by exponentiating and normalizing:  $\beta_{jk} = \exp(\eta_{jk}) / \sum_i^W \exp(\eta_{jk}^{(i)})$ , where the denominator sums over the vocabulary.
- **Generate regions:** for each region  $j < J$ ,
  - Draw the spatial mean  $\nu_j$  from a normal distribution.
  - Draw the precision matrix  $\Lambda_j$  from a Wishart distribution.
- Draw the distribution over regions  $\vartheta$  from a symmetric Dirichlet prior,  $\vartheta \sim \text{Dir}(\alpha_\vartheta \mathbf{1})$ .
- **Generate text and locations:** for each document  $d$ ,
  - Draw topic proportions from a symmetric Dirichlet prior,  $\theta \sim \text{Dir}(\alpha \mathbf{1})$ .
  - Draw the region  $r$  from the multinomial distribution  $\vartheta$ .
  - Draw the location  $y$  from the bivariate Gaussian,  $y \sim \mathcal{N}(\nu_r, \Lambda_r)$ .
  - For each word token,
    - \* Draw the topic indicator  $z \sim \theta$ .
    - \* Draw the word token  $w \sim \beta_{rz}$ .

#### 4.2.4 Inference

We apply mean-field variational inference: a fully-factored variational distribution  $Q$  is chosen to minimize the Kullback-Leibler divergence from the true distribution. Mean-field variational inference with conjugate priors is described in detail elsewhere (Bishop, 2006; Wainwright and Jordan, 2008); we restrict our focus to the issues that are unique to the geographic topic model.

We place variational distributions over all latent variables of interest:  $\theta, z, r, \vartheta, \eta, \mu, \sigma^2, \nu$ , and  $\Lambda$ , updating each of these distributions in turn, until convergence. The variational distributions over  $\theta$  and  $\vartheta$  are Dirichlet, and have closed form updates: each can be set to the sum of the expected counts, plus a term from the prior (Blei et al., 2003). The variational distributions  $q(z)$  and  $q(r)$  are categorical, and can be set proportional to the expected joint likelihood—to set  $q(z)$  we marginalize over  $r$ , and vice versa.<sup>7</sup> The updates for the multivariate Gaussian spatial parameters  $\nu$  and  $\Lambda$  are described by Penny (2001).

#### Regional Word Distributions

The variational region-topic distribution  $\eta_{jk}$  is normal, with uniform diagonal covariance for tractability. Throughout we will write  $\langle x \rangle$  to indicate the expectation of  $x$  under the variational distribution  $Q$ . Thus, the vector mean of the distribution  $q(\eta_{jk})$  is written  $\langle \eta_{jk} \rangle$ , while the variance (uniform across  $i$ ) of  $q(\eta)$  is written  $\mathcal{V}(\eta_{jk})$ .

To update the mean parameter  $\langle \eta_{jk} \rangle$ , we maximize the contribution to the variational bound  $L$  from the relevant terms:

$$L_{[\langle \eta_{jk}^{(i)} \rangle]} = \langle \log p(\mathbf{w} | \boldsymbol{\beta}, \mathbf{z}, \mathbf{r}) \rangle + \langle \log p(\eta_{jk}^{(i)} | \mu_k^{(i)}, \sigma_k^2) \rangle, \quad (4.1)$$

with the first term representing the likelihood of the observed words (recall that  $\boldsymbol{\beta}$  is computed deterministically from  $\eta$ ) and the second term corresponding to the prior. The likelihood term requires the expectation  $\langle \log \beta \rangle$ , but this is somewhat complicated by the normalizer  $\sum_i^W \exp(\eta^{(i)})$ , which sums over all terms in the vocabulary. As in previous work on logistic normal topic models, we use a Taylor approximation for this term (Blei and Lafferty, 2006a).

The prior on  $\eta$  is normal, so the contribution from the second term of the objective (Equation 4.1) is  $-\frac{1}{2\langle \sigma_k^2 \rangle} \langle (\eta_{jk}^{(i)} - \mu_k^{(i)})^2 \rangle$ . We introduce the following notation for expected counts:  $N(i, j, k)$  indicates the expected count of term  $i$  in region  $j$  and topic  $k$ , and  $N(j, k) = \sum_i N(i, j, k)$ . After some calculus, we can write the gradient  $\partial L / \partial \langle \eta_{jk}^{(i)} \rangle$  as

$$N(i, j, k) - N(j, k) \langle \beta_{jk}^{(i)} \rangle - \langle \sigma_k^{-2} \rangle (\langle \eta_{jk}^{(i)} \rangle - \langle \mu_k^{(i)} \rangle), \quad (4.2)$$

which has an intuitive interpretation. The first two terms represent the difference in expected counts for term  $i$  under the variational distributions  $q(z, r)$  and  $q(z, r, \beta)$ : this difference goes to zero when  $\beta_{jk}^{(i)}$  perfectly matches  $N(i, j, k)/N(j, k)$ . The third term penalizes  $\eta_{jk}^{(i)}$  for deviating from its prior  $\mu_k^{(i)}$ , but this penalty is proportional to the expected inverse variance  $\langle \sigma_k^{-2} \rangle$ . We apply gradient ascent to maximize the objective  $L$ . A similar set of calculations gives the gradient for the variance of  $\eta$ ; these are described in an online appendix (Eisenstein, 2010).

---

<sup>7</sup>Thanks to the mean field assumption, we can marginalize over  $z$  by first decomposing across all  $N_d$  words and then summing over  $q(z)$ .

## Base Topics

The base topic parameters are  $\mu_k$  and  $\sigma_k^2$ ; in the variational distribution,  $q(\mu_k)$  is normally distributed and  $q(\sigma_k^2)$  is Gamma distributed. Note that  $\mu_k$  and  $\sigma_k^2$  affect only the regional word distributions  $\eta_{jk}$ . An advantage of the logistic normal is that the variational parameters over  $\mu_k$  are available in closed form,

$$\begin{aligned}\langle \mu_k^{(i)} \rangle &= \frac{b^2 \sum_j^J \langle \eta_{jk}^{(i)} \rangle + \langle \sigma_k^2 \rangle a^{(i)}}{b^2 J + \langle \sigma_k^2 \rangle} \\ \mathcal{V}(\mu_k) &= (b^{-2} + J \langle \sigma_k^{-2} \rangle)^{-1},\end{aligned}$$

where  $J$  indicates the number of regions. The expectation of the base topic  $\mu$  incorporates the prior and the average of the generated region-topics—these two components are weighted respectively by the expected variance of the region-topics  $\langle \sigma_k^2 \rangle$  and the prior topical variance  $b^2$ . The posterior variance  $\mathcal{V}(\mu)$  is a harmonic combination of the prior variance  $b^2$  and the expected variance of the region topics.

The variational distribution over the region-topic variance  $\sigma_k^2$  has Gamma parameters. These parameters cannot be updated in closed form, so gradient ascent optimization is again required. The derivation of these updates is more involved; see appendix.

### 4.2.5 Implementation

Variational scheduling and initialization are important aspects of any hierarchical generative model, and are often under-discussed. In our implementation, the variational updates are scheduled as follows: given expected counts, we iteratively update the variational parameters on the region-topics  $\eta$  and the base topics  $\mu$ , until convergence. We then update the geographical parameters  $\nu$  and  $\Lambda$ , as well as the distribution over regions  $\vartheta$ . Finally, for each document we iteratively update the variational parameters over  $\theta$ ,  $z$ , and  $r$  until convergence, obtaining expected counts that are used in the next iteration of updates for the topics and their regional variants. We iterate an outer loop over the entire set of updates until convergence.

We initialize the model in a piecewise fashion. First we train a Dirichlet process mixture model on the locations  $y$ , using variational inference on the truncated stick-breaking approximation (Blei and Jordan, 2006). This automatically selects the number of regions  $J$ , and gives a distribution over each region indicator  $r_d$  from geographical information alone. We then run standard latent Dirichlet allocation to obtain estimates of  $z$  for each token (ignoring the locations). From this initialization we can compute the first set of expected counts, which are used to obtain initial estimates of all parameters needed to begin variational inference in the full model.

The prior  $a$  is the expected mean of each topic  $\mu$ ; for each term  $i$ , we set  $a^{(i)} = \log N(i) - \log N$ , where  $N(i)$  is the total count of  $i$  in the corpus and  $N = \sum_i N(i)$ . The variance prior  $b^2$  is set to 1, and the prior on  $\sigma^2$  is the Gamma distribution  $\mathcal{G}(2, 200)$ , encouraging minimal deviation from the base topics. The symmetric Dirichlet prior on  $\theta$  is set to  $\frac{1}{2}$ , and the symmetric Dirichlet parameter on  $\vartheta$  is updated from weak hyperpriors (Minka, 2003). Finally, the geographical model takes priors that are linked to the data: for each region, the mean is very weakly encouraged to be near the overall mean, and the covariance prior is set by the average covariance of clusters obtained by running  $K$ -means.

## 4.2.6 Evaluation

For a quantitative evaluation of the estimated relationship between text and geography, we assess our model’s ability to predict the geographic location of unlabeled authors based on their text alone.<sup>8</sup> This task may also be practically relevant as a step toward applications for recommending local businesses or social connections. A randomly-chosen 60% of authors are used for training, 20% for development, and the remaining 20% for final evaluation.

### Systems

We compare several approaches for predicting author location; we divide these into latent variable generative models and discriminative approaches.

**Geographic Topic Model** This is the full version of our system, as described in this paper. To predict the unseen location  $y_d$ , we iterate until convergence on the variational updates for the hidden topics  $z_d$ , the topic proportions  $\theta_d$ , and the region  $r_d$ . From  $r_d$ , the location can be estimated as  $\hat{y}_d = \arg \max_y \sum_j p(y|\nu_j, \Lambda_j)q(r_d = j)$ . The development set is used to tune the number of topics and to select the best of multiple random initializations.

**Mixture of Unigrams** A core premise of our approach is that modeling topical variation will improve our ability to understand geographical variation. We test this idea by fixing  $K = 1$ , running our system with only a single topic. This is equivalent to a Bayesian mixture of unigrams in which each author is assigned a single, regional unigram language model that generates all of his or her text. The development set is used to select the best of multiple random initializations.

**Supervised Latent Dirichlet Allocation** In a more subtle version of the mixture-of-unigrams model, we model each author as an admixture of regions. Thus, the latent variable attached to each author is no longer an index, but rather a vector on the simplex. This model is equivalent to supervised latent Dirichlet allocation (Blei and McAuliffe, 2008): each topic is associated with equivariant Gaussian distributions over the latitude and longitude, and these topics must explain both the text and the observed geographical locations. For unlabeled authors, we estimate latitude and longitude by estimating the topic proportions and then applying the learned geographical distributions. This is a linear prediction

$$f(\bar{z}_d; \mathbf{a}) = (\bar{z}_d^\top \mathbf{a}^{\text{lat}}, \bar{z}_d^\top \mathbf{a}^{\text{lon}})$$

for an author’s topic proportions  $\bar{z}_d$  and topic-geography weights  $\mathbf{a} \in \mathbb{R}^{2K}$ .

**Text Regression** We perform linear regression to discriminatively learn the relationship between words and locations. Using term frequency features  $x_d$  for each author, we predict locations with word-geography weights  $\mathbf{a} \in \mathbb{R}^{2W}$ :

$$f(x_d; \mathbf{a}) = (x_d^\top \mathbf{a}^{\text{lat}}, x_d^\top \mathbf{a}^{\text{lon}})$$

---

<sup>8</sup>Alternatively, one might evaluate the attributed regional memberships of the words themselves. While the Dictionary of American Regional English (Cassidy and Hall, 1985) attempts a comprehensive list of all regionally-affiliated terms, it is based on interviews conducted from 1965-1970, and the final volume (covering Si-Z) is not yet complete.

System	Regression		Classification accuracy (%)	
	Mean Dist. (km)	Median Dist. (km)	Region (4-way)	State (49-way)
Geographic topic model	<b>900</b>	<b>494</b>	<b>58</b>	24
Mixture of unigrams	947	644	53	19
Supervised LDA	1055	728	39	4
Text regression	948	712	41	4
$K$ -nearest neighbors	1077	853	37	2
Mean location	1148	1018		
Most common class			37	<b>27</b>

Table 4.1: Location prediction results; lower scores are better on the regression task, higher scores are better on the classification task. Distances are in kilometers. Mean location and most common class are computed from the test set. Both the geographic topic model and supervised LDA use the best number of topics from the development set (10 and 5, respectively).

Weights are trained to minimize the sum of squared Euclidean distances, subject to  $L_1$  regularization:

$$\begin{aligned} & \sum_d (\mathbf{x}_d^\top \mathbf{a}^{\text{lat}} - y_d^{\text{lat}})^2 + (\mathbf{x}_d^\top \mathbf{a}^{\text{lon}} - y_d^{\text{lon}})^2 \\ & + \lambda_{\text{lat}} \|\mathbf{a}^{\text{lat}}\|_1 + \lambda_{\text{lon}} \|\mathbf{a}^{\text{lon}}\|_1 \end{aligned}$$

The minimization problem decouples into two separate latitude and longitude models, which we fit using the `glmnet` elastic net regularized regression package (Friedman et al., 2010) which has obtained good results on other text-based prediction tasks (Joshi et al., 2010). Regularization parameters were tuned on the development set. The  $L_1$  penalty outperformed  $L_2$  and mixtures of  $L_1$  and  $L_2$ .

Note that for both word-level linear regression here, and the topic-level linear regression in SLDA, the choice of squared Euclidean distance dovetails with our use of spatial Gaussian likelihoods in the geographic topic models, since optimizing  $\mathbf{a}$  is equivalent to maximum likelihood estimation under the assumption that locations are drawn from equivariant circular Gaussians centered around each  $f(\mathbf{x}_d; \mathbf{a})$  linear prediction. We experimented with decorrelating the location dimensions by projecting  $\mathbf{y}_d$  into the principal component space, but this did not help text regression.

**$K$ -Nearest Neighbors** Linear regression is a poor model for the multimodal density of human populations. As an alternative baseline, we applied supervised  $K$ -nearest neighbors to predict the location  $\mathbf{y}_d$  as the average of the positions of the  $K$  most similar authors in the training set. We computed term-frequency inverse-document frequency features and applied cosine similarity over their first 30 principal components to find the neighbors. The choices of principal components, IDF weighting, and neighborhood size  $K = 20$  were tuned on the development set.

## Metrics

Our principle error metrics are the mean and median distance between the predicted and true location in kilometers.<sup>9</sup> Because the distance error may be difficult to interpret, we also report

<sup>9</sup>For convenience, model training and prediction use latitude and longitude as a naively projected 2D Euclidean space. However, properly measuring the physical distance between points on the Earth's surface requires calculating

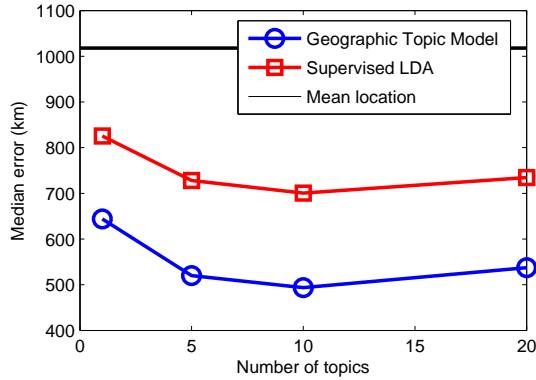


Figure 4.2: The effect of varying the number of topics on the median regression error (lower is better).

accuracy of classification by state and by region of the United States. Our data includes the 48 contiguous states plus the District of Columbia; the U.S. Census Bureau divides these states into four regions: West, Midwest, Northeast, and South.<sup>10</sup> Note that while major population centers straddle several state lines, most region boundaries are far from the largest cities, resulting in a clearer analysis.

## Results

As shown in Table 4.1, the geographic topic model achieves the strongest performance on all metrics. All differences in performance between systems are statistically significant ( $p < .01$ ) using the Wilcoxon-Mann-Whitney test for regression error and the  $\chi^2$  test for classification accuracy. Figure 4.2 shows how performance changes as the number of topics varies.

Note that the geographic topic model and the mixture of unigrams use identical code and parametrization – the only difference is that the geographic topic model accounts for topical variation, while the mixture of unigrams sets  $K = 1$ . These results validate our basic premise that it is important to model the interaction between topical and geographical variation.

Text regression and supervised LDA perform especially poorly on the classification metric. Both methods make predictions that are averaged across each word in the document: in text regression, each word is directly multiplied by a feature weight; in supervised LDA the word is associated with a latent topic first, and then multiplied by a weight. For these models, all words exert an influence on the predicted location, so uninformative words will draw the prediction towards the center of the map. This yields reasonable distance errors but poor classification accuracy. We hypothesized that  $K$ -nearest neighbors would be a better fit for this metric, but its performance is poor at all values of  $K$ . Of course it is always possible to optimize classification accuracy directly, but such an approach would be incapable of predicting the exact geographical location, which is the focus of our evaluation (given that the desired geographical partition

---

the great circle distance, for which we use the Haversine formula (Sinnott, 1984). For the continental U.S., degree-space is a reasonable approximation for modeling. While the dimensions are are differently sized—in the geographical center of the U.S., latitude is about 111 km/degree while longitude is about 85 km/degree—since our model includes spatial covariance, it effectively stretches and rotates as much as necessary. Extending the model to a continental scale would require a more sophisticated approach.

<sup>10</sup>[http://www.census.gov/geo/www/us\\_regdiv.pdf](http://www.census.gov/geo/www/us_regdiv.pdf)

	<b>"basketball"</b> PISTONS KOBE LAKERS game DUKE NBA CAVS STUCKEY JETS KNICKS	<b>"popular music"</b> album music beats artist video #LAKERS iTUNES tour produced vol	<b>"daily life"</b> tonight shop weekend getting going chilling ready discount waiting iam	<b>"emoticons"</b> :) haha :d :( ;) :p xd :/ hahaha hahah	<b>"chit chat"</b> lol smh jk yea wyd coo ima wassup somethin jp
Boston		CELTICS victory BOSTON CHARLOTTE	playing daughter PEARL alive war comp	BOSTON	:p gna loveee <i>ese</i> exam suttin sippin
N. California		THUNDER KINGS GIANTS pimp trees clap	SIMON dl mountain seee	6am OAKLAND	<i>pues</i> hella koo SAN fckn hella flirt hut iono OAKLAND
New York		NETS KNICKS	BRONX	iam cab	oww wasssup nm
Los Angeles		#KOBE #LAKERS AUSTIN	#LAKERS load HOLLYWOOD imm MICKEY TUPAC	omw tacos hr HOLLYWOOD	af <i>papi</i> raining th bomb coo HOLLYWOOD wyd coo af <i>nada</i> tacos messin fasho bomb
Lake Erie		CAVS CLEVELAND OHIO BUCKS od COLUMBUS	premiere prod joint TORONTO onto designer CANADA village burr	stink CHIPOTLE tipsy	foul WIZ salty excuses lames officer lastnight

Table 4.2: Example base topics (top line) and regional variants. For the base topics, terms are ranked by log-odds compared to the background distribution. The regional variants show words that are strong compared to both the base topic and the background. Foreign-language words are shown in *italics*, while terms that are usually in proper nouns are shown in **SMALL CAPS**. See Table 4.3 for definitions of slang terms; see Section 4.2.7 for more explanation and details on the methodology.

is unknown). Note that the geographic topic model is also not trained to optimize classification accuracy.

#### 4.2.7 Analysis

Our model permits analysis of geographical variation in the context of topics that help to clarify the significance of geographically-salient terms. Table 4.2 shows a subset of the results of one randomly-initialized run, including five hand-chosen topics (of 50 total) and five regions (of 13, as chosen automatically during initialization). Terms were selected by log-odds comparison. For the base topics we show the ten strongest terms in each topic as compared to the background word distribution. For the regional variants, we show terms that are strong both regionally and topically: specifically, we select terms that are in the top 100 compared to both the background distribution and to the base topic. The names for the topics and regions were chosen by the authors.

Nearly all of the terms in column 1 (“**basketball**”) refer to sports teams, athletes, and place names—encouragingly, terms tend to appear in the regions where their referents reside. Column

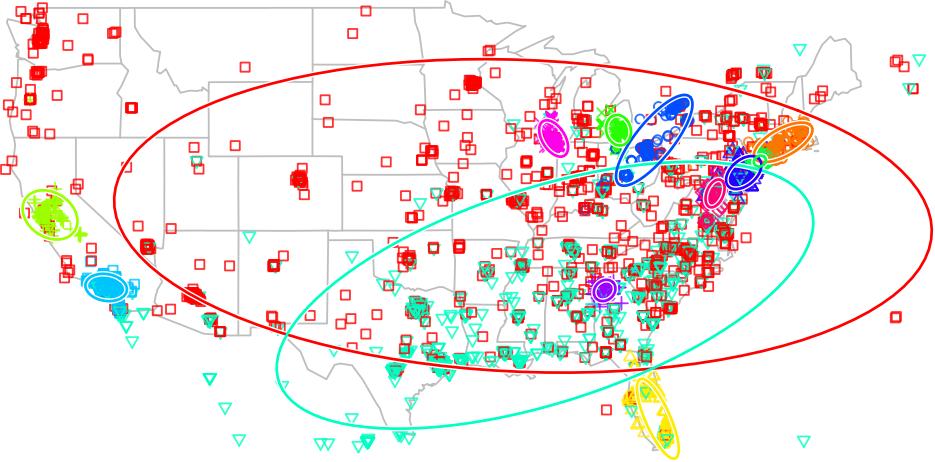


Figure 4.3: Regional clustering of the training set obtained by one randomly-initialized run of the geographical topic model. Each point represents one author, and each shape/color combination represents the most likely cluster assignment. Ellipses represent the regions' spatial means and covariances. The same model and coloring are shown in Table 4.2.

2 contains several proper nouns, mostly referring to popular music figures (including **PEARL** from the band Pearl Jam).<sup>11</sup> Columns 3–5 are more conversational. Spanish-language terms (*papi*, *pues*, *nada*, *ese*) tend to appear in regions with large Spanish-speaking populations—it is also telling that these terms appear in topics with emoticons and slang abbreviations, which may transcend linguistic barriers. Other terms refer to people or subjects that may be especially relevant in certain regions: *tacos* appears in the southern California region and *cab* in the New York region; **TUPAC** refers to a rap musician from Los Angeles, and **WIZ** refers to a rap musician from Pittsburgh, not far from the center of the “Lake Erie” region.

A large number of slang terms are found to have strong regional biases, suggesting that slang may depend on geography more than standard English does. The terms *af* and *hella* display especially strong regional affinities, appearing in the regional variants of multiple topics (see Table 4.3 for definitions). While research in perceptual dialectology does confirm the link of *hella* to Northern California (Bucholtz et al., 2007), we caution that our findings are merely suggestive, and a more extensive analysis must be undertaken before making definitive statements about the regional membership of individual terms. We view the geographic topic model as an exploratory tool that may be used to facilitate such investigations.

Figure 4.3 shows the regional clustering on the training set obtained by one run of the model. Each point represents an author, and the ellipses represent the bivariate Gaussians for each region. There are nine compact regions for major metropolitan areas, two slightly larger regions that encompass Florida and the area around Lake Erie, and two large regions that partition the country roughly into north and south.

---

<sup>11</sup>This analysis is from an earlier version of our dataset that contained some Twitterbots, including one from a Boston-area radio station. The bots were removed for the evaluation in Section 4.2.6, though the numerical results are nearly identical.

term	definition	term	definition
af	as fuck (very)	jk	just kidding
coo	cool	jp	just playing (kidding)
dl	download	koo	cool
fasho	for sure	lol	laugh out loud
gna	going to	nm	nothing much
hella	very	od	overdone (very)
hr	hour	omw	on my way
iam	I am	smh	shake my head
ima	I'm going to	suttin	something
imm	I'm	wassup	what's up
iono	I don't know	wyd	what are you doing?
lames	lame (not cool) people		

Table 4.3: A glossary of non-standard terms from Table 4.2. Definitions are obtained by manually inspecting the context in which the terms appear, and by consulting [www.urbandictionary.com](http://www.urbandictionary.com).

#### 4.2.8 Related Work

The relationship between language and geography has been a topic of interest to linguists since the nineteenth century (Johnstone, 2010). An early work of particular relevance is Kurath’s *Word Geography of the Eastern United States* (1949), in which he conducted interviews and then mapped the occurrence of equivalent word pairs such as *stoop* and *porch*. The essence of this approach—identifying variable pairs and measuring their frequencies—remains a dominant methodology in both dialectology (Labov et al., 2006) and sociolinguistics (Tagliamonte, 2006). Within this paradigm, computational techniques are often applied to post hoc analysis: logistic regression (Sankoff et al., 2005) and mixed-effects models (Johnson, 2009) are used to measure the contribution of individual variables, while hierarchical clustering and multidimensional scaling enable aggregated inference across multiple variables (Nerbonne, 2009). However, in all such work it is assumed that the relevant linguistic variables have already been identified—a time-consuming process involving considerable linguistic expertise. We view our work as complementary to this tradition: we work directly from raw text, identifying both the relevant features and coherent linguistic communities.

An active recent literature concerns geotagged information on the web, such as search queries (Backstrom et al., 2008) and tagged images (Crandall et al., 2009). This research identifies the geographic distribution of individual queries and tags, but does not attempt to induce any structural organization of either the text or geographical space, which is the focus of our research. More relevant is the work of Mei et al. (2006), in which the distribution over latent topics in blog posts is conditioned on the geographical location of the author. This is somewhat similar to the supervised LDA model that we consider, but their approach assumes that a partitioning of geographical space into regions is already given. The dataset we released with the published version of this paper has been used for several studies investigating the geolocation task. Wing and Baldridge (2011); Roller et al. (2012) focus just on the geolocation task, Eisenstein et al. (2011b) present a variant of the model here, and Hong et al. (2012); Ahmed et al. (2013) also discover geographical topics using different models.

Methodologically, our cascading topic model is designed to capture multiple dimensions of variability: topics and geography. Mei et al. (2007) include sentiment as a second dimension in a

topic model, using a switching variable so that individual word tokens may be selected from either the topic or the sentiment. However, our hypothesis is that individual word tokens reflect both the topic and the geographical aspect. Sharing this intuition, Paul and Girju (2010) build topic-aspect models for the cross product of topics and aspects. They do not impose any regularity across multiple aspects of the same topic, so this approach may not scale when the number of aspects is large (they consider only two aspects). We address this issue using cascading distributions; when the observed data for a given region-topic pair is low, the model falls back to the base topic. The use of cascading logistic normal distributions in topic models follows earlier work on dynamic topic models (Blei and Lafferty, 2006b; Xing, 2005). Paul and Dredze (2012) present an alternative Dirichlet approach to combining multiple factors into topic-word distributions. (There has also been previous work with hierarchical Dirichlets and their generalizations for n-gram models of differing prefix lengths (MacKay and Peto, 1995; Teh, 2006) and interpolating against document-level topics (Wallach, 2008, §3.4).) Eisenstein et al. (2011b) extend the additive cascading topics approach here to have sparse deviations, and Roberts et al. (2013) embed them in a more general covariate-influenced topic model.

## 4.3 Demographic lexical variation through U.S. Census data

(This section was originally published as O'Connor et al. (2010b).)

### 4.3.1 Introduction

In this study, we apply the model of §4.2 to discover demographic language variation from text and metadata, in order to explain both linguistic variation and demographic features through a set of generative distributions, each of which is associated with a (latent) community of speakers. Thus, our model is capable of discovering both the relevant sociolinguistic communities, as well as the key dimensions of lexical variation.

### 4.3.2 Data

We start with the same dataset used in Section 4.2, focusing on a randomly-selected subset of 4875 authors. For this research, we have extended the corpus with detailed demographic metadata. While it is difficult to identify the demographic attributes of individual speakers, we can cross-reference speaker locations against U.S. Census data to extract aggregate demographic statistics of each user's geographic location.<sup>12</sup> We use the Zip Code Tabulation Areas (ZCTA) level of granularity, which partitions the U.S. into 33,178 geographic features (typically polygons). Using a standard geospatial tool,<sup>13</sup> we match each author's location to the area that contains it, and use the area's demographics as that author's demographic metadata. The set of features that we consider are shown in Table 4.4. To give a rough view of word association, for each variable we show the top five words ranked by the sample-corrected average demographic value among authors who use them at least once.<sup>14</sup> Some terms are telling by themselves, though some aspects are obviously issues with the relatively small sample and narrow timeframe (e.g. *mangoville*); based in part on this experience, in Chapter 5 we focus on longer-term data.

---

<sup>12</sup>We use year 2000 data from Summary Files 1 and 3: <http://www.census.gov/support/cen2000.html>

<sup>13</sup>PostGIS: <http://postgis.net/>

<sup>14</sup>We rank by the lower bound of the 95% confidence interval for the mean:  $\hat{\mu} - 1.96 \hat{\sigma} / \sqrt{n}$ . Using the raw average always places rare words in the top ranks, which is often due to statistical noise.

Demographic Variable	Mean	SD	Words with Highest Average
Percent White	52.1%	29.0	leno, fantastic, holy, military, review
Percent African-American	32.2%	29.1	lml, momz, midterms, bmore, fuccin
Percent Hispanic	15.7%	18.3	cuando, estoy, pero, eso, gracias
Percent English speakers	73.7%	18.4	#lowkey, porter, #ilovefamu, nc, atlanta
Percent population in urban areas	95.1%	14.3	odeee, thatt, m2, madd, mangoville
Percent family households	64.1%	14.4	mangoville, lightskin, iin, af, aha
Median annual income†	\$39,045	(26k, 59k)	mangoville, tuck, itunes, jim, dose

Figure 4.4: List of demographic variables used, selected from 2000 U.S. Census data, along with their mean and standard deviation among authors in the data, and the words with the highest sample-adjusted average values. The procedure for selecting words is described in Section 4.3.2; some analysis appears in Section 4.3.4. (Income is shown in dollars, but the model uses log-dollars. The SD column shows  $(\hat{\mu} \pm \hat{\sigma})$  computed on the log scale.)

While geographical aggregate statistics are frequently used to proxy for individual socioeconomic status in research fields such as public health (e.g., Rushton, 2008), it is clear that interpretation must proceed with caution. Consider an author from a zip code in which 60% of the residents are Hispanic:<sup>15</sup> we do not know the likelihood that the author is Hispanic, because the set of Twitter users is not a representative sample of the overall population. Polling research suggests that users of both Twitter (Smith and Rainie, 2010) and geolocation services (Zickuhr and Smith, 2010) are much more diverse with respect to age, gender, race and ethnicity than the general population of Internet users. Nonetheless, at present we can only use aggregate statistics to make inferences about the geographic communities in which our authors live, and not the authors themselves.

### 4.3.3 Model

Our latent variable model combines demographic metadata with microblog text. The goal is to extract a set of latent sociolinguistic communities which are coherent with respect to both data sources. Our model combines a multinomial distribution over text with a multivariate Gaussian over the demographic statistics; these generative components are unified in a Dirichlet process mixture, in which each speaker has a latent “community” index. The model corresponds to the single-membership, “mixture of unigrams” version of Section 4.2’s Geographic Topic Model, where  $K = 1$ . For clarity, we give a complete description of this version of the model as follows.

We hypothesize a generative stochastic process that produces the text and the demographic data for each author. This generative process includes a set of latent variables; we will recover a variational distribution over these latent variables using a mean field representation. The plate diagram for this model is shown in Figure 4.5. The key latent variable is the community membership of each author, which we write  $c_d$ ; this variable selects a distribution over both the metadata and the text. Each distribution over metadata is a multivariate Gaussian with parameters  $\mu$  and  $\Lambda$ ; each distribution over text is a multinomial with parameter  $\beta$ . Overall, we can describe the generative process as:

- Draw the community proportions  $\vartheta$  from a stick-breaking prior,
- **Generate the community distributions.** For each community  $i$ ,

<sup>15</sup>In the U.S. Census, the official ethnonym is *Hispanic or Latino*; for brevity we use *Hispanic* in the rest of this thesis.

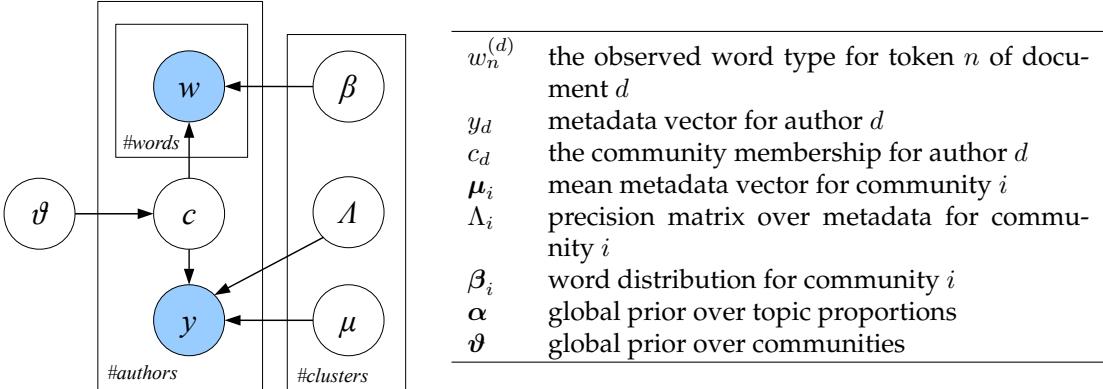


Figure 4.5: Plate diagram for our model of text and demographics, with a table of all random variables. The document indices in the figure are implicit, as are the priors on  $\mu$ ,  $\Lambda$ ,  $\beta$ .

- Draw the metadata mean  $\mu_i$  from a multivariate Normal prior,
- Draw the metadata precision matrix  $\Lambda_i$  from a Wishart prior,
- Draw the word distribution  $\beta_i$  from a Dirichlet prior,
- **Generate the text and metadata.** For each author  $d$ ,
  - Draw the community  $c_d$  from the distribution  $\vartheta$ ,
  - Draw the metadata  $y_d$  from a Gaussian with mean  $\mu_{c_d}$  and precision  $\Lambda_{c_d}$ ,
  - Draw the bag of words  $w_d$  from the multinomial  $\beta_{c_d}$ ,

We apply mean field variational inference to recover a posterior distribution over the random variables in this model (Bishop, 2006), as described in §4.2.4. For efficiency, we initialized by running the Dirichlet process mixture model on the demographic data alone. The number of active clusters from this initialization was used as the fixed number of clusters when running the full model on both text and demographics data.

#### 4.3.4 Analysis

Figure 4.4 shows the words associated with large values for each demographic feature, ranked by sample-corrected averages (described above). This is informative and should be viewed as an exploratory method in its own right; however, correlations between demographic variables make it difficult to disentangle the underlying relationships between demography and lexical frequencies.

In contrast, Figure 4.3.3 summarizes all of the sociolinguistic clusters identified by our model. All clusters are shown. Each row shows a cluster that corresponds to a distribution over demographic information, along with a set of characteristic terms as chosen by likelihood ratio. This cluster analysis allows us to associate each term with a complete demographic profile.

Several of the top terms refer to subjects which attracted only an ephemeral interest, and would likely not appear in a dataset taken from a longer timespan. The term *19th* refers to an event on March 19, 2010 that was a frequent subject of conversation in this dataset. The term *olive* usually refers to the Olive Garden restaurant; *mangoville* refers to a restaurant in New York City. Florida

	Mean Vector	Top Words
Cluster 1		rsvp, ent, guest, blvd, broadcasting, details, bash, lls, retweet, —, #free, —, hosting, pow, vibe, 31, vol, —, feat, 2nite
Cluster 2		en, de, el, que, es, por, se, un, los, pero, una, como, para, lo, del, te, si, eso, la, tu
Cluster 3		ii, dha, yu, uu, yuu, dhat, lols, lolss, lml, qo, qot, w—, myy, iim, qet, yuh, smhh, niqqa, buh, &&
Cluster 4		#ilovefamu, lmbo, grind, official, awards, #lowkey, #famusextape, jake, track, spirit, #thatissall, mental, famu, praying, studying, you're, bible, midterm, joy, awesome
Cluster 5		lls, jawn, neighbors, joints, nivea, #famusextape, sextape, #epicfail, cuddle, broad, midterms, jeezy, #thatissall, basic, nigga, waited, tmobile, menu, bcuz, famu
Cluster 6		gimmie, hosted, —, download, b-day, dl, limit, drake, mix, dj, mc, salute, //, #unotfromthehoodif, ft, exclusive, birthday, models, -, lab
Cluster 7		:], ;], --, ^, bahaha, :d, papi, ^_^, ily, aha, =], fck, ha-hah, lovee, ew, yess, :/, #urparentsever, mangoville, jamaica
Cluster 8		dats, dat, dis, wat, da, watz, dey, wats, den, gud, wen, gravity, niggaz, jus, der, fuk, rite, dem, tha, dese
Cluster 9		rare, 19th, olive, simply, adam, agent, coffee, obama, awesome, 400, hockey, leno, thomas, worked, pentagon, #fb, tone, presents, larry, peppers

Figure 4.6: Demographic mean vectors and most salient words per cluster. Demographic variables are shown on a normalized scale; zero indicates the population mean, and the axis tick marks denote  $\pm 1$  standard deviation; see Table 4.4 for their values. For each cluster  $k$ , words shown are the top-20 most highly ranked by the ratio of topic probability against background probability:  $\frac{\beta_k[w]}{\tilde{p}(w)}$ .

A&M, a historically-black university, appears in the terms *#ilovefamu* and *#famusextape*. Terms that start with hashtags (e.g., *#epicfail*, *#thatisall*) often represent “trends” that are shown on all users’ Twitter pages; many users participate by adding their own commentary on such tags (Kwak et al., 2010). The topic lists also contain several names, including *leno*, *obama*, and the musicians *drake* and *jeezy*.

One cluster (number 2) contains exclusively Spanish words. These words exhibit a strong mutual association, as many authors will use only Spanish words and few if any words in English. This cluster is relatively diffuse with regard to demographic data, and thus we would not expect it to be detected without recourse to the linguistic properties of the speakers. Note that while this cluster contains a high proportion of Hispanics, it also appears to contain an above-average number of white speakers. We see two potential explanations: the speakers in this cluster may come from mixed white-Hispanic neighborhoods, or the individual authors may identify as both ethnicities.

We see a number of phenomena which are characteristic of computer-mediated communication, including emoticons, phonetic spelling, and abbreviations (Tagliamonte and Denis, 2008). Emoticons (e.g., :) are grouped in cluster 7, which contains many Hispanics and is above-average with respect to income. Phonetic spelling is used in clusters 3 and 8, which are the two lowest-income clusters and contain the fewest whites. The other group with an above-average number of blacks is cluster 5, and the associated language is somewhat more standardized. Relative to clusters 3 and 8, cluster 5 is wealthier, more urban, and contains more whites and fewer Hispanics.

Clusters 1, 4, and 9 are substantially less urban than the other clusters, and they contain the most standard English words. Of particular interest is cluster 4, which is the only non-urban cluster with lower-than-average income; the fact that this cluster is still relatively standard may provide hint to the relative importance of urbanity and wealth with respect to relative frequency of standard and “vernacular” language. These findings are better thought of as provisional, given the complexities and limitations of the data, such as age and demographic skew of Twitter versus the full population the Census considers, and the small timeframe of the dataset.

## 4.4 Conclusion and reflections on textual social data analysis

The relationship between language, geography, and social identity has traditionally been studied with respect to micro-level phenomena, such as phonological features or individual words, that have been manually identified by researchers. In this work, we take a more holistic approach to discovering groups of words whose variation is associated with social factors, using a generative model that operates on authors’ microblog texts. In §4.2 our model extracts geographic-linguistic communities that are coherent with respect to geographic location and text, while in §4.3 we discover sociolinguistic communities that are coherent with respect to both demographic metadata and text. This model also identifies individual terms that are especially characteristic of these communities in social media.

The overwhelming prevalence of non-standard and novel words was a surprise to us, and made evident that new forms of language are apparently being invented in the medium of online discussions. While the one-week dataset was convenient for getting started with analysis and model development, it is apparent that many aspects of language in social media might be ephemeral, and the results in this chapter are sometimes very specific to the dates the dataset was gathered. One lesson is that, in order to collect data for general linguistic analysis, it is important to use data from over a long timespan; we do this in support of a part-of-speech tagging system in Owoputi et al. (2013) (not included in this thesis). But also, this suggests that the temporal as-

pect of how geography, demographics, and language interact—that is, the social determinants of language change—is interesting in itself. This is the focus of the next chapter.

Finally, within the range of computational methods pursued in this thesis, this work stands at a particular position: it consists of a fairly complex, latent variable model used for exploratory analysis. The mixed-membership clustering is very useful for understanding broader groups of terms and authors, and it helped us get a sense that there exist large and widely used vocabularies—entire linguistic worlds—that are very divergent from standard English.

However, latent variable approaches have some practical issues for exploratory data analysis (EDA). For one, on large datasets, they can be computationally intensive to run, which runs counter to the need of rapid iterative data analytic experimentation when doing EDA. Newer research since this work has found ways to improve inference speed for topic models (such as stochastic variational approximations (Hoffman et al., 2013), or more specialized moment-based methods (Anandkumar et al., 2012; Arora et al., 2012)), but given the intractability of either posterior or maximum likelihood inference, runtimes may always be much slower than simpler count-based methods, such as the pointwise mutual information approach described in Chapter 2. In fact, after finishing the published version of Section 4.2, we experimented further with spatial density plots of individual terms, yielding fascinating patterns shown in the next chapter (Figure 5.1), and further investigated there. The experience of doing direct analysis of words against spatial coordinates also informed the *MiTextExplorer* tool, which attempts to better automate and make more interactive this type of exploratory text data analysis (Chapter 2).

Also, in work outside this thesis, we found PMI methods useful to summarize geographically-specific terms in Chinese social media (Bamman et al., 2012), to better understand the nature of government censorship of that medium. There, regions were pre-defined by the data (for example, some are provinces in China), and thus are a discrete variable amenable to measuring PMI against word frequencies. If there is access to coordinate-level location data, the geographic topic model developed in this chapter is more powerful in some respects because it discovers spatial regions, instead of pre-defining them; but this can also make analysis more difficult, since it requires effort to understand and verify that the inferred regions are meaningful for a particular analysis.

Still, latent variable models have many uses. The next chapter utilizes latent variables not for grouping terms, but for understanding temporal and diffusion dynamics of individual words. And Chapter 6, which investigates textually described events between political actors, makes extensive use of a topic model in order to infer latent classes of events.

## *Chapter 5*

---

# **Social determinants of linguistic diffusion in social media**

---

(An earlier version of this work was published as Eisenstein et al. (2012).)

## **5.1 Synopsis**

Computer-mediated communication is driving fundamental changes in the nature of written language. We investigate these changes by statistical analysis of a dataset comprising 107 million Twitter messages (authored by 2.7 million unique user accounts). Using a latent vector autoregressive model to aggregate across thousands of words, we identify high-level patterns in diffusion of linguistic change over the United States. Our model is robust to unpredictable changes in Twitter's sampling rate, and provides a probabilistic characterization of the relationship of macro-scale linguistic influence to a set of demographic and geographic predictors. The results of this analysis offer support for prior arguments that focus on geographical proximity and population size. However, demographic similarity – especially with regard to race – plays an even more central role, as cities with similar racial demographics are far more likely to share linguistic influence. Rather than moving towards a single unified “netspeak” dialect, language evolution in computer-mediated communication reproduces existing fault lines in spoken American English.

## **5.2 Introduction**

An increasing proportion of informal communication is conducted in written form, mediated by technology such as smartphones and social media platforms. Written language has been forced to adapt to meet the demands of synchronous conversation, resulting in a creative burst of new forms, such as emoticons, abbreviations, phonetic spellings, and other neologisms (Androustosopoulos, 2000; Anis, 2007; Herring, 2012). Such changes have often been considered as a single, uniform dialect — both by researchers (Crystal, 2006; Squires, 2010) and throughout the popular press (Thurlow, 2006; Squires, 2010). But despite the fact that social media facilitates instant communication between distant corners of the earth, the adoption of new written forms is often sharply delineated by geography and demographics (Eisenstein et al., 2010, 2011c; Schwartz et al., 2013). For example: the abbreviation *ikr* (*I know, right?*) occurs six times more frequently in the Detroit area than in the United States overall; the emoticon ^-^ occurs four times more frequently

in Southern California; the phonetic spelling *suttin* (*something*) occurs five times more frequently in New York City.

These differences raise questions about how language change spreads in online communication. What groups are influential, and which communities evolve together? Is written language moving toward global standardization or increased fragmentation? As language is a crucial constituent of personal and group identity, examination of the competing social factors that drive language change can shed new light on the hidden structures that shape our society. This chapter offers a new technique for inducing networks of linguistic influence and co-evolution from raw word counts. We then seek explanations for this network in a set of demographic and geographic predictors.

A wave of recent research has shown how social media datasets can enable large-scale analysis of patterns of communication (Lotan et al., 2011; Wu et al., 2011), sentiment (Dodds et al., 2011; Thelwall, 2009; Mitchell et al., 2013), and influence (Lazer et al., 2009; Aral and Walker, 2012; Bond et al., 2012; Gomez-Rodriguez et al., 2012; Bakshy et al., 2012). Such work has generally focused on tracking the spread of discrete behaviors, such as using a piece of software (Aral and Walker, 2012), reposting duplicate or near-duplicate content (Leskovec et al., 2009; Cha et al., 2010; Lotan et al., 2011), voting in political elections (Bond et al., 2012), or posting a hyperlink to online content (Gomez-Rodriguez et al., 2012; Bakshy et al., 2012). Tracking linguistic changes poses a significant additional challenge, as we are concerned not with the first appearance of a word, but with the bursts and lulls in its popularity over time (Altmann et al., 2009). In addition, the well known “long-tail” nature of both word counts and city sizes (Zipf, 1949/2012) ensures that most counts for words and locations will be sparse, rendering simple frequency-based methods inadequate.

Language change has long been an active area of research, and a variety of theoretical models have been proposed. In the **wave** model, linguistic innovations spread through interactions over the course of an individual’s life, so the movement of linguistic innovation from one region to another depends on the density of interactions (Bailey, 1973). In the simplest version of this model, the probability of contact between two individuals depends on their distance, so linguistic innovations should diffuse continuously through space. The **gravity** and **cascade** models refine this view, arguing that the likelihood of contact between individuals from two cities depends on the size of the cities as well as their distance; thus, linguistic innovations should be expected to travel between large cities first (Trudgill, 1974; Labov, 2003). However, Nerbonne and Heeringa find little evidence that population size impacts diffusion of pronunciation differences in dialects of the Netherlands (Nerbonne and Heeringa, 2007).

Cultural factors also play an important role in both the diffusion of, and resistance to, language change. Many words and phrases have entered the standard English lexicon from minority dialects (Lee, 1999); conversely, there is evidence that minority groups in the United States resist regional sound changes associated with European American speakers (Gordon, 2000), and that racial differences in speech persist even in conditions of very frequent social contact (Rickford, 1985). At present there are few quantitative sociolinguistic accounts of how geography and demographics interact; nor are their competing roles explained in the menagerie of theoretical models of language change, such as evolutionary biology (Zhang and Gong, 2013; Baxter et al., 2006), dynamical systems (Niyogi and Berwick, 1997), Nash equilibria (Trapa and Nowak, 2000), Bayesian learners (Reali and Griffiths, 2010), and agent-based simulations (Fagyal et al., 2010). In general, such research is concerned with demonstrating that a proposed theoretical framework can account for observed phenomena like geographical distribution of linguistic features and their rate of adoption over time. In contrast, this chapter is concerned with fitting a model to a large corpus of text data from individual language users, and analyzing the social meaning of the resulting

parameters.

Research on reconstructing language phylogenies from cognate tables is more closely related (Gray and Atkinson, 2003; Gray et al., 2009; Bouckaert et al., 2012; Dunn et al., 2011), but rather than a phylogenetic process in which languages separate and then develop in relative independence, we have closely-related varieties of a single language, which are in constant interaction. Other researchers have linked databases of typological linguistic features (such as morphological complexity) with geographical and social properties of the languages' speech communities (Lupyan and Dale, 2010; Daumé III, 2009). Again, our interest is in more subtle differences within the same language, rather than differences across the entire set of world languages. The typological atlases and cognate tables that are the basis such work are inapplicable to our problem, requiring us to take a corpus-based approach (Szmrecsanyi, 2011), estimating an influence network directly from raw text.

The overall aim of this work is to build a computational model capable of identifying the demographic and geographic factors that drive the spread of newly popular words in online text. To this end, we construct a statistical procedure for recovering networks of linguistic diffusion from raw word counts, even as the underlying social media sampling rate changes unaccountably. We present a procedure for Bayesian inference in this model, capturing uncertainty about the induced diffusion network. We then consider a range of demographic and geographic factors which might explain the networks induced from this model, using a logistic regression analysis. This lends support to prior work on the importance of population and geography, but reveals a strong role for racial homophily at the level of city-to-city linguistic influence.

### 5.3 Materials and methods

We conducted a statistical analysis of a corpus of public data from the microblog site Twitter, from 2009–2012. The corpus includes 107 million messages, mainly in English, from more than 2.7 million unique user accounts. Each message contains GPS coordinates to locations in the continental United States. The data was temporally aggregated into 165 week-long bins. After taking measures to remove marketing-oriented accounts, each user account was associated with one of the 200 largest Metropolitan Statistical Areas (MSA) in the United States, based on their geographical coordinates. The 2010 United Census provides detailed demographics for MSAs. By linking this census data to changes in word frequencies, we can obtain an aggregate picture of the role of demographics in the diffusion of linguistic change in social media.

Empirical research suggests that Twitter's user base is younger, more urban, and more heavily composed of ethnic minorities, in comparison with the overall United States population (Mislove et al., 2011; Duggan and Smith, 2013). Our analysis does *not* assume that Twitter users are a representative demographic sample of their geographic areas. Rather, we assume that on a macro scale, the diffusion of words between metropolitan areas depends on the overall demographic properties of those areas, and not on the demographic properties specific to the Twitter users that those areas contain. Alternatively, the use of population-level census statistics can be justified on the assumption that the demographic skew introduced by Twitter — for example, towards younger individuals — is approximately homogeneous across cities. Table 5.1 shows the average demographics for the 200 MSAs considered in our study.

Linguistically, our analysis begins with the 100,000 most frequent terms overall. We narrow this list to 4,854 terms whose frequency changed significantly over time. The excluded terms have little dynamic range; they would therefore not substantially affect the model parameters, but would increase the computational cost if included. We then manually further refine the list to

	mean	st. dev
Population	1,170,000	2,020,000
Log Population	13.4	0.9
% Urbanized	77.1	12.9
Median Income	61,800	11,400
Log Median Income	11.0	0.2
Median age	36.8	3.9
% Renter	34.3	5.2
% Af. Am	12.9	10.6
% Hispanic	15.0	17.2

Table 5.1: Statistics of metropolitan statistical areas. Mean and standard deviation for demographic attributes of the 200 Metropolitan Statistical Areas (MSAs) considered in this study, from 2010 Census data.

2,603 English words, by excluding names, hashtags, and foreign language terms. Both a complete list of terms and more detailed procedures for data acquisition are given in §5.6.

Figure 5.1 shows the geographical distribution of six words over time. The first row shows the word *ion*, which is a shortened form of *I don't*, as in *ion even care*. Systematically coding a random sample of 300 occurrences of the string *ion* in our dataset revealed two cases of the traditional chemistry sense of *ion*, and 294 cases that clearly matched *I don't*. This word displays increasing popularity over time, but remains strongly associated with the Southeast. In contrast, the second row shows the emoticon -\_- (indicating annoyance), which spreads from its initial bases in coastal cities to nationwide popularity. The third row shows the abbreviation *ctfu*, which stands for *cracking the fuck up* (i.e., laughter). At the beginning of the sample it is active mainly in the Cleveland area; by the end, it is widely used in Pennsylvania and the mid-Atlantic, but remains rare in the large cities to the west of Cleveland, such as Detroit and Chicago. What explains the non-uniform spread of this term's popularity?

While individual examples are intriguing, we seek an aggregated account of the spatiotemporal dynamics across many words, which we can correlate against geographic and demographic properties of metropolitan areas. Due to the complexity of drawing inferences about influence and demographics from raw word counts, we perform this process in stages. (A block diagram of the procedure is shown in Figure 5.2.) First, we model word frequencies as a dynamical system, using Bayesian inference over the latent spatiotemporal activation of each word. We use sequential Monte Carlo (Godsill et al., 2004) to approximate the distribution over spatiotemporal activations with a set of samples. Within each sample, we induce a model of the linguistic dynamics between metropolitan areas, which we then discretize into a set of pathways. Finally, we perform logistic regression to identify the geographic and demographic factors that correlate with the induced linguistic pathways. By aggregating across samples, we can estimate the confidence intervals of the resulting logistic regression parameters.

### 5.3.1 Modeling spatiotemporal lexical dynamics in social media data

This section describes our approach for modeling lexical dynamics in our data. We represent our data as counts  $c_{w,r,t}$ , which is the number of individuals who used the word  $w$  at least once in MSA  $r$  at time  $t$  (i.e., one week). (Mathematical notation is summarized in Table 5.2. We do not consider the total number of times a word is used, since there are many cases of a single

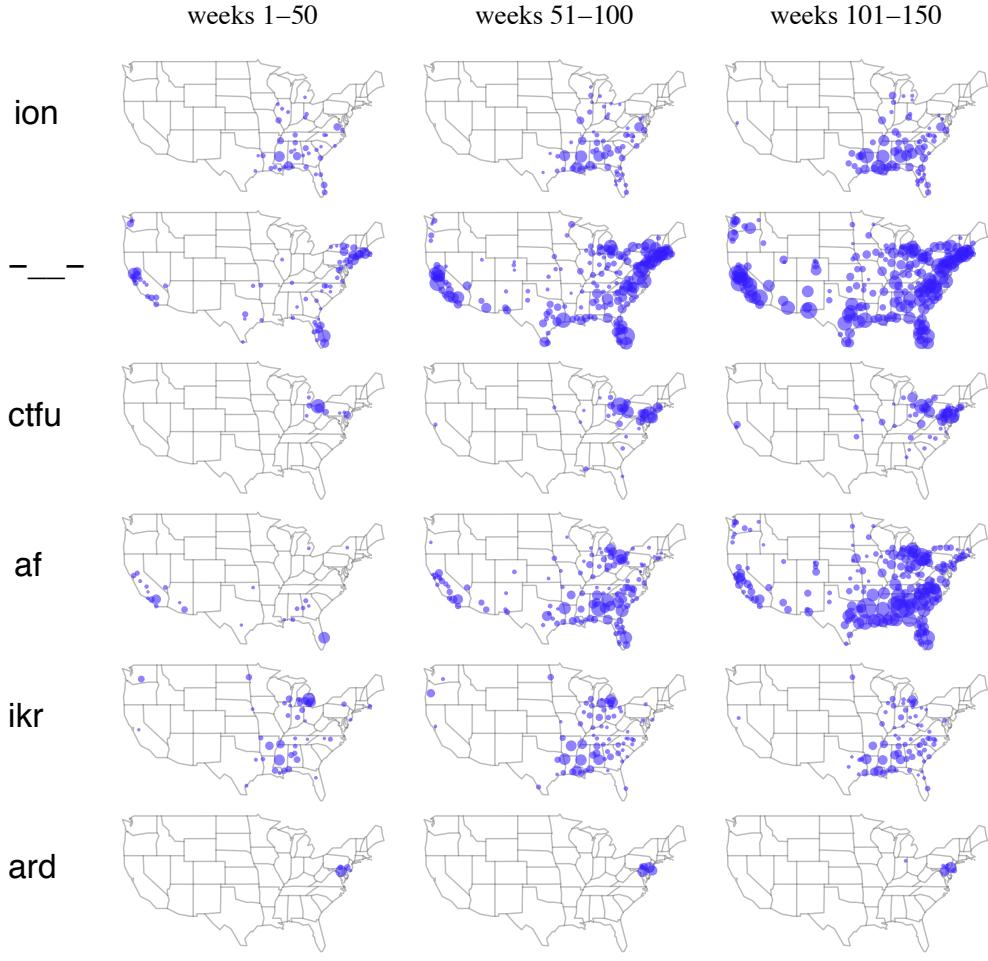


Figure 5.1: Change in frequency for six words: *ion*, *---*, *ctfu*, *af*, *ikr*, *ard*. Blue circles indicate cities where on average, at least 0.1% of users use the word during a week. A circle's area is proportional to the word's probability.

individual using a single word hundreds or thousands of times.) To capture the dynamics of these counts, we employ a latent vector autoregressive model, based on the binomial distribution with a logistic link function. The use of latent variable modeling is motivated by properties of the data that are problematic for simpler autoregressive models that operate directly on word counts and frequencies (without a latent variable). We begin by briefly summarizing these problems; we then present our model, describe the details of inference and estimation, and offer some examples of the inferences that our model supports.

### Challenges for direct autoregressive models

The simplest modeling approach would be an autoregressive model that operates directly on the word counts or frequencies (Wei, 1994). A major challenge for such models is that Twitter offers only a sample of all public messages, and the sampling rate can change in unclear ways (Morstatter et al., 2013). For example, for much of the timespan of our data, Twitter's documentation

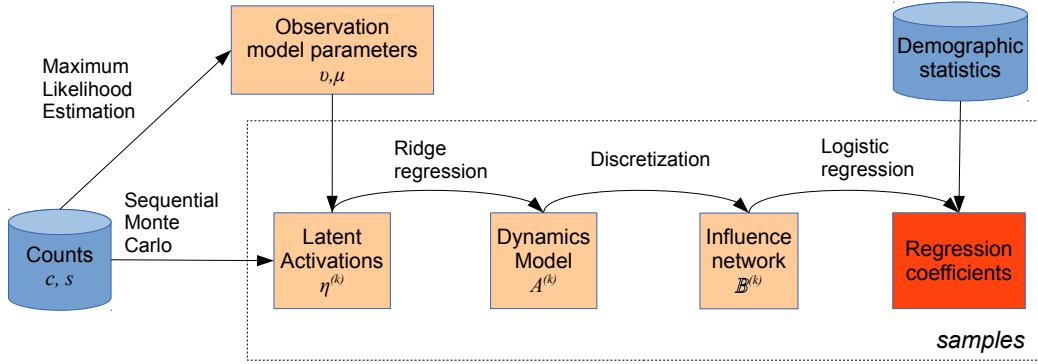


Figure 5.2: Block diagram for our statistical modeling procedure. The dotted outline indicates repetition across samples drawn from sequential Monte Carlo.

$c_{w,r,t}$	Number of individuals who used word $w$ in metropolitan area $r$ during week $t$ .
$s_{r,t}$	Number of individuals who posted messages in metropolitan area $r$ at time $t$ .
$p_{w,r,t}$	Empirical probability that an individual from metropolitan area $r$ will use word $w$ during week $t$ .
$\eta_{w,r,t}$	Latent spatiotemporal activation for word $w$ in metropolitan area $r$ at time $t$ .
$\nu_{w,t}$	Global activation for word $w$ at time $t$ .
$\mu_{r,t}$	Regional activation (“verbosity”) for metropolitan area $r$ at time $t$ .
$a_{r_1,r_2}$	Autoregressive coefficient from metropolis $r_1$ to $r_2$ .
$A = \{a_{r_1,r_2}\}$	Complete autoregressive dynamics matrix.
$\sigma_{w,r}^2$	Autoregressive variance for $\eta_{w,r,t}$ , for all times $t$ .
$\lambda$	Variance of zero-mean Gaussian prior over each $a_{r_1,r_2}$ .
$\omega_{w,r,t}^{(k)}$	Weight of sequential Monte Carlo hypothesis $k$ for word $w$ , metropolis $r$ , and time $t$ .
$z_{r_1,r_2}$	$z$ -score of $a_{r_1,r_2}$ , computed from empirical distribution over Monte Carlo samples.
$\bar{\mathbb{B}}$	Set of ordered city pairs for whom $a_{r_1,r_2}$ is significantly greater than zero, computed over all samples.
$\mathbb{B}^{(k)}$	Top $L$ ordered city pairs, as sorted by the bottom of the 95% confidence interval on $\{a_{r_1,r_2}^{(k)}\}$ .
$Q$	Random distribution over discrete networks, designed so that the marginal frequencies for “sender” and “receiver” metropolises are identical to their empirical frequencies in the model-inferred network.

Table 5.2: Table of mathematical notation.

implies that the sampling rate is approximately 10%; but in 2010 and earlier, the sampling rate appears to be 15% or 5%.<sup>1</sup> After 2010, the volume growth in our data is relatively smooth, implying that the sampling is fair (unlike the findings of Morstatter et al., which focus on a more problematic case involving query filters, which we do not use).

Raw counts are not appropriate for analysis, because the MSAs have wildly divergent numbers of users and messages. New York City has four times as many active users as the 10th largest MSA

<sup>1</sup>This estimate is based on inspection of message IDs modulo 100, which appears to be how sampling was implemented at that time.

(San Francisco-Oakland, CA), twenty times as many as the 50th largest MSA (Oklahoma City, OK), and 200 times as many as the 200th largest MSA (Yakima, WA); these ratios are substantially larger when we count messages instead of active users. This necessitates normalizing the counts to frequencies  $p_{w,r,t} = c_{w,r,t}/s_{r,t}$ , where  $s_{r,t}$  is the number of individuals who have written at least one message in region  $r$  at time  $t$ . The resulting frequency  $p_{w,r,t}$  is the empirical probability that a random user in  $(r, t)$  used the word  $w$ . Word frequencies treat large and small cities more equally, but suffer from several problems:

- The frequency  $p_{w,r,t}$  is *not* invariant to a change in the sampling rate: if, say, half the messages are removed, the probability of seeing a user use any particular word goes down, because  $s_{r,t}$  will decrease more slowly than  $c_{w,r,t}$  for any  $w$ . The changes to the global sampling rate in our data drastically impact  $p_{w,r,t}$ .
- Users in different cities can be more or less actively engaged with Twitter: for example, the average New Yorker contributed 55 messages to our dataset, while the average user within the San Francisco-Oakland MSA contributed 21 messages. Most cities fall somewhere in between these extremes, but again, this “*verbosity*” may change over time.
- Word popularities can be driven by short-lived global phenomena, such as holidays or events in popular culture (e.g., TV shows, movie releases), which are not interesting from the perspective of persistent changes to the lexicon. We manually removed terms that directly refer to such events (as described in Section 5.6), but there may be unpredictable second-order phenomena, such as an emphasis on words related to outdoor cooking and beach trips during the summer, and complaints about boredom during the school year.
- Due to the long-tail nature of both word counts and city populations (Clauset et al., 2009), many word counts in many cities are zero at any given point in time. This floor effect means that least squares models, such as Pearson correlations or the Kalman smoother, are poorly suited for this data, in either the  $c_{w,r,t}$  or  $p_{w,r,t}$  representations.

## A latent vector autoregressive model

To address these issues, we build a latent variable model that controls for these confounding effects, yielding a better view of the underlying frequency dynamics for each word. Instead of working with raw frequencies  $p_{w,r,t}$ , we perform inference over latent variables  $\eta_{w,r,t}$ , which represent the underlying *activation* of word  $w$  in MSA  $r$  at time  $t$ . This latent variable parameterizes a distribution for the count data  $c_{w,r,t}$  via a binomial distribution with the number of trials  $s_{r,t}$ . A binomial distribution requires a frequency parameter, which we attain by passing  $\eta$  through the logistic function, where  $\text{Logistic}(\eta) = 1/(1 + e^{-\eta})$ .

An  $\eta$ -only model, therefore, would be

$$c_{w,r,t} \sim \text{Binomial}(s_{r,t}, \text{Logistic}(\eta_{w,r,t})) \quad (5.1)$$

This is a very simple generalized linear model with a logit link function (Gelman and Hill, 2006), in which the maximum likelihood estimate of  $\eta$  would simply be a log-odds reparameterization of the probability of a user using the word,  $\hat{\eta}_{w,r,t} = \log(p_{w,r,t}/(1 - p_{w,r,t}))$ . By itself, this model corresponds to directly using  $p_{w,r,t}$ , and has all the same problems as noted in the previous section; in addition, the estimate  $\hat{\eta}_{w,r,t}$  goes to negative infinity when  $c_{w,r,t} = 0$ .

The advantage of the logistic binomial parameterization is that it allows an additive combination of effects to control for confounds. The  $\eta$  variables still represent differences in log-odds, but

after controlling for “base rate” effects. To this end, we include two additional parameters  $\nu_{w,t}$  and  $\mu_{r,t}$ :

$$c_{w,r,t} \sim \text{Binomial}(s_{r,t}, \text{Logistic}(\eta_{w,r,t} + \nu_{w,t} + \mu_{r,t})). \quad (5.2)$$

The parameter  $\nu_{w,t}$  represents the overall activation of the word  $w$  at time  $t$ , thus accounting for non-geographical changes, such as when a word becomes more popular everywhere at once. The parameter  $\mu_{r,t}$  represents the “verbosity” of MSA  $r$  at time  $t$ , which varies for the reasons mentioned above. These parameters control for global effects due to  $t$ , such as changes to the API sampling rate. (Because  $\mu_{r,t}$  and  $\nu_{w,t}$  both interact with  $t$ , it is unnecessary to introduce a main effect for  $t$ .) These parameters subtract out nuisance effects, enabling a more stable estimate of  $\eta$ .

We can now measure lexical dynamics in terms of the latent variable  $\eta$  rather than the raw counts  $c$ . We take the simplest possible approach, modeling  $\eta$  as a first-order linear dynamical system with Gaussian noise (Gelb, 1974),

$$\eta_{w,r,t} \sim N\left(\sum_{r'} a_{r',r} \eta_{w,r',t-1}, \sigma_{w,r}^2\right). \quad (5.3)$$

The dynamics matrix  $A = \{a_{r_1,r_2}\}$  is shared over both words and time; we also assume homogeneity of variance within each metropolitan area (per word), using the variance parameter  $\sigma_{w,r}^2$ . These simplifying assumptions are taken to facilitate statistical inference, by keeping the number of parameters at a reasonable size. If it is possible to detect clear patterns of linguistic diffusion under this linear homoscedastic model, then more flexible models should show even stronger effects, if they can be estimated successfully. We leave this investigation for future work. It is important to observe that this model *does* differentiate directionality: in general,  $a_{r_1,r_2} \neq a_{r_2,r_1}$ . The coefficient  $a_{r_1,r_2}$  reflects the extent to which  $\eta_{r_1,t}$  predicts  $\eta_{r_2,t+1}$ , and vice versa for  $a_{r_2,r_1}$ . In the extreme case that  $r_1$  ignores  $r_2$ , while  $r_2$  imitates  $r_1$  perfectly, we will have  $a_{r_1,r_2} = 1$  and  $a_{r_2,r_1} = 0$ . Note that both coefficients can be positive, in the case that  $\eta_{r_1}$  and  $\eta_{r_2}$  evolve smoothly and synchronously; indeed, such mutual connections appear frequently in the induced networks.

Equation 5.2 specifies the *observation* model, and Equation 5.3 specifies the *dynamics* model; together, they specify the joint probability distribution,

$$P(\eta, c | s; A, \sigma^2, \mu, \nu) = P(c | \eta, s; \mu, \nu)P(\eta; A). \quad (5.4)$$

Because the observation model is non-Gaussian, the standard Kalman smoother cannot be applied. Inference under non-Gaussian distributions is often handled via second-order Taylor approximation, as in the extended Kalman filter (Gelb, 1974), but a second-order approximation to the Binomial distribution is unreliable when the counts are small. In contrast, sequential Monte Carlo sampling permits arbitrary parametric distributions for both the observations and system dynamics (Cappe et al., 2007). Forward-filtering backward sampling (Godsill et al., 2004) gives smoothed samples from the distribution  $P(\eta_{w,1:R,1:T} | c_{w,1:R,1:T}, s_{1:R,1:T}, A)$ , so for each word  $w$ , we obtain a set of sample trajectories  $\eta_{w,1:R,1:T}^{(k)}$  for  $k \in \{1, \dots, K = 100\}$ .

## Inference and estimation

The total dimension of  $\eta$  is equal to the product of the number of MSAs (200), words (2,603), and time steps (165), requiring inference over 85 million interrelated random variables. To facilitate inference and estimation, we adopt a stagewise procedure. First we make estimates of the parameters  $\nu$  (overall activation for each word) and  $\mu$  (region-specific verbosity), assuming  $\eta_{w,r,t} = 0, \forall w, r, t$ . Next, we perform inference over  $\eta$ , assuming a simplified dynamics matrix  $\tilde{A}$ ,

which is diagonal. Last, we perform inference over the full dynamics matrix  $A$ , from samples from this distribution. See Figure 5.2 for a block diagram of the inference and estimation procedure.

The parameters  $\nu$  (global word activation) and  $\mu$  (region-specific verbosity) are estimated first. We begin by computing a simplified  $\bar{\nu}_w$  as the inverse logistic function of the total frequency of word  $w$ , across all time steps. Next, we compute the maximum likelihood estimates of each  $\mu_{r,t}$  via gradient descent. We then hold  $\mu$  fixed, and compute the maximum likelihood estimates of each  $\nu_{w,t}$ . Inference over the latent spatiotemporal activations  $\eta_{w,r,t}$  is performed via Monte Carlo Expectation Maximization (MCEM) (Wei and Tanner, 1990). For each word  $w$ , we construct a diagonal dynamics matrix  $\tilde{A}_w$ . Given estimates of  $\tilde{A}_w$  and  $\sigma_w^2$ , we use the sequential Monte Carlo (SMC) algorithm of forward-filtering backward sampling (FFBS) (Godsill et al., 2004) to draw samples of  $\eta_{w,1:T}$ ; this constitutes the E-step of the MCEM process. Next, we apply maximum-likelihood estimation to update  $\tilde{A}_w$  and  $\sigma_w^2$ ; this constitutes the M-step. These updates are repeated until either the parameters converge or we reach a limit of twenty iterations. We describe each step in more detail:

- **E-step.** The E-step consists of drawing samples from the posterior distribution over  $\eta$ . FFBS appends a backward pass to any SMC filter that produces a set of hypotheses and weights  $\{\eta_{w,r,t}^{(k)}, \omega_{w,r,t}^{(k)}\}_{1 \leq k \leq K}$ . The role of the backward pass is to reduce variance by resampling the hypotheses according to the joint smoothing distribution. Our forward pass is a standard bootstrap filter (Cappe et al., 2007): by setting the proposal distribution  $q(\eta_{w,r,t} | \eta_{w,r,t-1})$  equal to the transition distribution  $P(\eta_{w,r,t} | \eta_{w,r,t-1}; \tilde{A}_w, \sigma_{w,r}^2)$ , the forward weights are equal to the recursive product of the observation likelihoods,

$$\omega_{w,r,t}^{(k)} = \omega_{w,r,t-1}^{(k)} P(c_{w,r,t} | \eta_{w,r,t}, s_{w,t}; \nu_{w,t}, \mu_{r,t}). \quad (5.5)$$

The backward pass uses these weights, and returns a set of unweighted hypotheses that are drawn directly from  $P(\eta_{w,r,t} | c_{w,r,t}, s_{r,t}; \nu_{w,t}, \mu_{r,t})$ . More complex SMC algorithms — such as resampling, annealing, and more accurate proposal distributions — did not achieve higher likelihood than the bootstrap filter.

- **M-step.** The M-step consists of computing the average of the maximum likelihood estimates of  $\tilde{A}_w$  and  $\sigma_w^2$ . Within each sample, maximum likelihood estimation is straightforward: the dynamics matrix  $\tilde{A}_w$  is obtained by least squares, and  $\sigma_{w,r}^2$  is set to the empirical variance  $\frac{1}{T} \sum_t^T (\eta_{w,r,t} - \tilde{a}_{w,r} \eta_{w,r,t-1})^2$ .

## Examples

Figure 5.3 shows the result of this modeling procedure for several example words. In the right panel, each sample of  $\eta$  is shown with a light dotted line. In the left panel, the empirical word frequencies are shown with circles, and the smoothed frequencies for each sample are shown with dotted lines. Large cities generally have a lower variance over samples, because the variance of the maximum *a posteriori* estimate of the binomial decreases with the total event count. For example, in Figure 5.3(c), the samples of  $\eta$  are tightly clustered for Philadelphia (the sixth-largest MSA in the United States), but are diffuse for Youngstown (the 95th largest MSA). Note also that the relationship between frequency and  $\eta$  is not monotonic — for example, the frequency of *ion* increases in Memphis over the duration of the sample, but the value of  $\eta$  decreases. This is because of the parameter for background word activation,  $\nu_{w,t}$ , which increases as the word attains more general popularity. The latent variable model is thus able to isolate MSA-specific activation from nuisance effects that include the overall word activation and Twitter’s changing sampling rate.

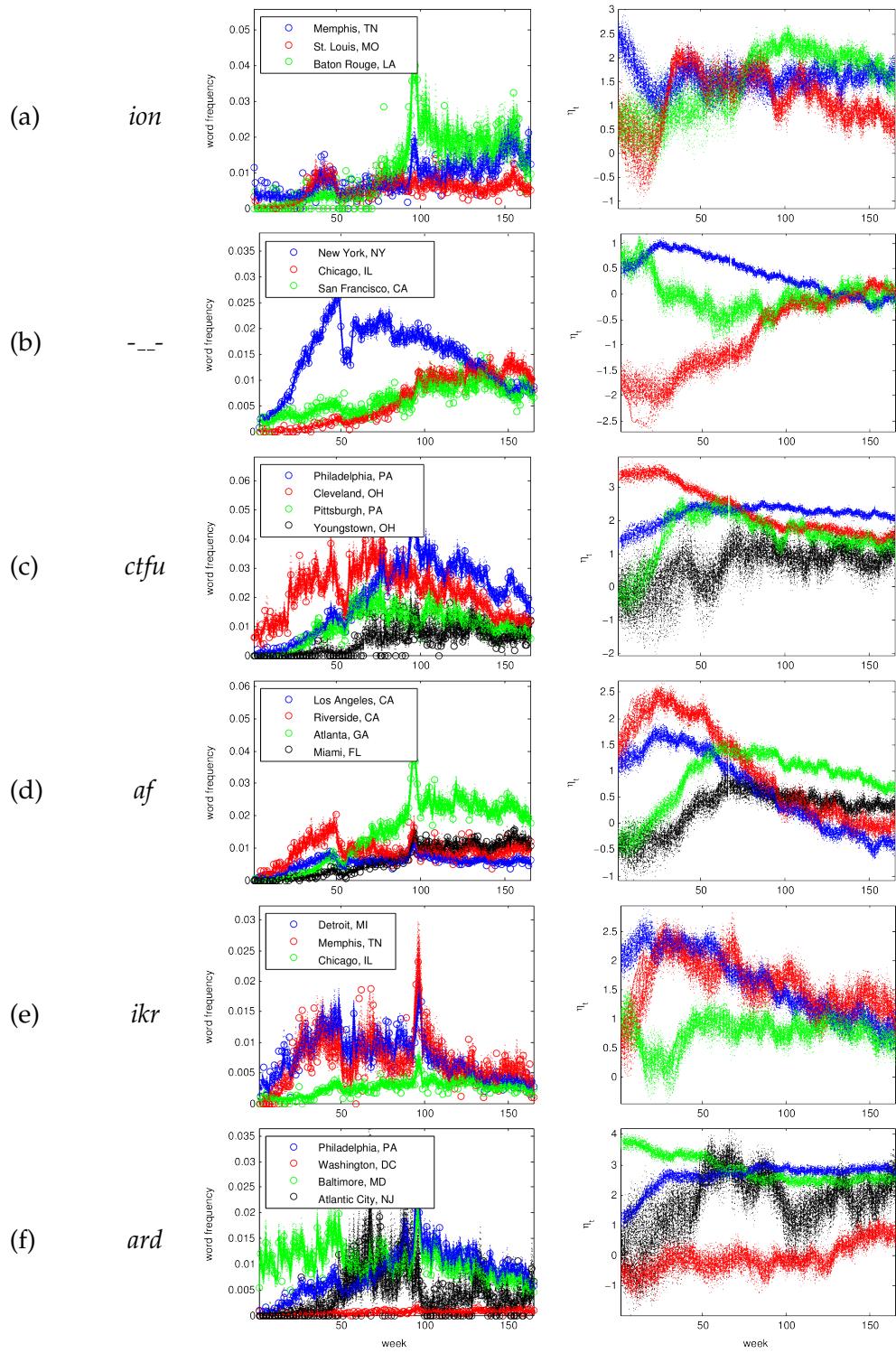


Figure 5.3: Left: empirical term frequencies (circles) and their Monte Carlo smoothed estimates (dotted lines); Right: Monte Carlo smoothed estimates of  $\eta_t$ .

### 5.3.2 Constructing a network of linguistic diffusion

Having obtained samples from the distribution  $P(\eta | c, s)$  over latent spatiotemporal activations, we now move to estimate the system dynamics, which describes the pathways of linguistic diffusion. Given the simple Gaussian form of the dynamics model (Equation 5.3), the coefficients  $A$  can be obtained by ordinary least squares. We perform this estimation separately within each of the  $K$  sequential Monte Carlo samples  $\eta^{(k)}$ , obtaining  $K$  dense matrices  $A^{(k)}$ , for  $k \in \{1, \dots, K\}$ .

The coefficients of  $A^{(k)}$  are not in meaningful units, and their relationship to demographics and geography will therefore be difficult to interpret, model, and validate. Instead, we prefer to use a binarized, *network* representation,  $\mathbb{B}$ . Given such a network, we can directly compare the properties of linked MSAs with the properties of randomly selected pairs of MSAs not in  $\mathbb{B}$ , offering face validation of the proposed link between macro-scale linguistic influence and the demographic and geographic features of cities.

Specifically, we are interested in a set of pairs of MSAs,  $\mathbb{B} = \{\langle r_1, r_2 \rangle\}$ , for which we are confident that  $a_{r_1, r_2} > 0$ , given the uncertainty inherent in estimation across sparse word counts. Monte Carlo inference enables this uncertainty to be easily quantified: we compute  $z$ -scores  $z_{r_1, r_2}$  for each ordered city pair, using the empirical mean and standard deviation of  $a_{r_1, r_2}^{(k)}$  across samples  $k \in \{1, \dots, K\}$ . We select pairs whose  $z$ -score exceeds a threshold  $z^{(\text{thresh})}$ , denoting the selected set  $\bar{\mathbb{B}} = \{\langle r_i, r_j \rangle : z_{i,j} > z^{(\text{thresh})}\}$ . To compute uncertainty around a large number of coefficients, we apply the Benjamini-Hochberg False Discovery Rate (FDR) correction for multiple hypothesis testing (Benjamini and Hochberg, 1995), which controls the expected proportion of false positives in  $\bar{\mathbb{B}}$  as

$$\text{FDR}(z^{(\text{thresh})}) = \frac{P_{\text{null}}(z_{i,j} > z^{(\text{thresh})})}{\tilde{P}(z_{i,j} > z^{(\text{thresh})})} = \frac{1 - \Phi(z^{(\text{thresh})})}{[R(R-1)]^{-1} \sum_{i \neq j} 1\{z_{i,j} > z^{(\text{thresh})}\}}, \quad (5.6)$$

where the null probability is a one-sided hypothesis that  $z$  exceeds  $z^{(\text{thresh})}$  under a standard normal distribution, which we would expect if  $a_{i,j}$  values were random; this has probability  $1 - \Phi(z^{(\text{thresh})})$ , where  $\Phi$  is the Gaussian CDF.  $\tilde{P}$  is the simulation-generated empirical distribution over  $z(a_{i,j})$  values. If high  $z$ -scores occur much more often under the model ( $\tilde{P}$ ) than we would expect by chance ( $P_{\text{null}}$ ), only a small proportion should be expected to be false positives; the Benjamini-Hochberg ratio is an upper bound on the expected proportion of false positives in  $\bar{\mathbb{B}}$ . To obtain  $\text{FDR} < 0.05$ , the individual test threshold is approximately  $z^{(\text{thresh})} = 3.2$ , or in terms of  $p$ -values,  $p < 6 \times 10^{-4}$ . We see 510 dynamics coefficients survive this threshold; these indicate high-probability pathways of linguistic diffusion. The associated set of city pairs is denoted  $\bar{\mathbb{B}}_{0.05}$ .

Figure 5.4 shows a sparser network  $\bar{\mathbb{B}}_{0.001}$ , induced using a more stringent threshold of  $\text{FDR} < 0.001$ . The role of geography is apparent from the figure: there are dense connections within regions such as the Northeast, Midwest, and West Coast, and relatively few cross-country connections. For example, we observe many connections among the West Coast cities of San Diego, Los Angeles, San Jose, San Francisco, Portland, and Seattle (from bottom to top on the left side of the map), but few connections from these cities to other parts of the country.

**Practical details.** To avoid overfitting and degeneracy in the estimation of  $A^{(k)}$ , we place a zero-mean Gaussian prior on each element  $a_{r_1, r_2}^{(k)}$ , tuning the variance  $\lambda$  by grid search on the log-likelihood of a held-out subset of time slices within  $\eta_{1:T}$ . The maximum *a posteriori* estimate of  $A$  can be computed in closed form via ridge regression. Lags of length greater than one are accounted for by regressing the values of  $\eta_t$  against the moving average from the previous ten time steps. Results without this smoothing are broadly similar.

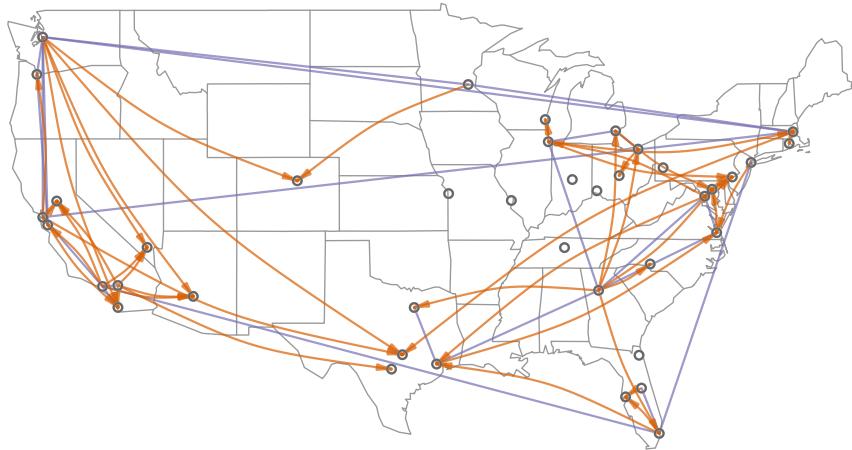


Figure 5.4: Induced network, showing significant coefficients among the 40 most populous MSAs (using an FDR < 0.001 threshold, yielding 254 links). Blue edges represent bidirectional influence, when there are directed edges in both directions; orange links are unidirectional.

### 5.3.3 Geographic and demographic correlates of linguistic diffusion

By analyzing the properties of pairs of metropolitan areas that are connected in the network  $\mathbb{B}$ , we can quantify the geographic and demographic drivers of online language change. Specifically, we construct a logistic regression to identify the factors that are associated with whether a pair of cities have a strong linguistic connection. The positive examples are pairs of MSAs with strong transmission coefficients  $a_{r_1, r_2}$ ; an equal number of negative examples is sampled randomly from a distribution  $Q$ , which is designed to maintain the same empirical distribution of MSAs that appears in the positive examples. This ensures that each MSA appears with roughly the same frequency in the positive and negative pairs, eliminating a potential confound.

The independent variables in this logistic regression include geographic and demographic properties of pairs of MSAs. We include the following demographic attributes: median age, log median income, and the proportions of, respectively, African Americans, Hispanics, individuals who live in urbanized areas, and individuals who rent their homes. The proportion of European Americans was omitted because of a strong negative correlation with the proportion of African Americans; the proportion of Asian Americans was omitted because it is very low for the overwhelming majority of the 200 largest MSAs. These raw attributes are then converted into both asymmetric and symmetric predictors, using the raw difference and its absolute value. The symmetric predictors indicate pairs of cities that are likely to share influence; besides the demographic attributes, we include geographical distance and, as a control, the log of the sum of populations. The asymmetric predictors are properties that may make an MSA likely to be the driver of online language change. Besides the raw differences of the six demographic attributes, we include the log difference in population. For a given demographic attribute, a negative regression coefficient for the absolute difference would indicate that similarity is important; a positive regression coefficient for the (asymmetric) raw difference would indicate that regions with large values of this attribute tend to be senders rather than receivers of linguistic innovations. All variables are standardized.

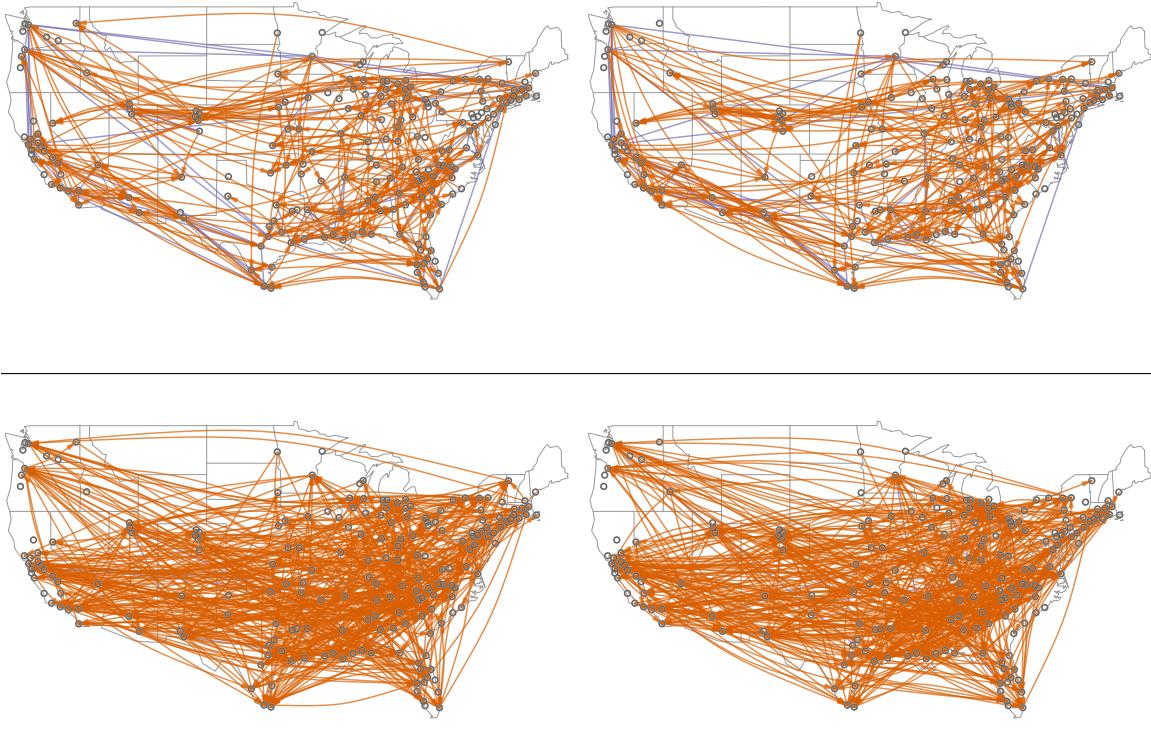


Figure 5.5: Top: two sample networks inferred by the model. (Unlike Figure 5.4, all 200 cities are shown.) Bottom: two “negative” networks, sampled from  $Q$ ; these are samples from the non-linked pair distribution  $Q$ , which is constructed to have the same marginal distributions over senders and receivers as in the inferred network. A blue line indicates directed edges in both directions between the pair of cities; orange lines are unidirectional.

To visually verify the geographic distance properties of our model, Figure 5.5 compares networks obtained by discretizing  $A^{(k)}$  against networks of randomly-selected MSA pairs, sampled from  $Q$ . Histograms of these distances are shown in Figure 5.6, and their average values are shown in Table 5.3. The networks induced by our model have many more short-distance connections as compared to chance. Table 5.3 also shows that many other demographic attributes are more similar among cities that are linked in our model’s network.

A logistic regression can show the extent to which each of the above predictors relates to the dependent variable, the binarized linguistic influence. However, the posterior uncertainty of the estimates of the logistic regression coefficients depends not only on the number of instances (MSA pairs), but principally on the variance in the Monte Carlo-based estimates for  $A^{(k)}$ , which in turn depends on the sampling variance and the size of the observed spatiotemporal word counts. To properly account for this complex variance, we run the logistic regression separately within each Monte Carlo sample  $k$ , and report the empirical standard errors of the logistic coefficients across the samples.

**Practical details** This procedure requires us to discretize the dynamics network *within* each sample, which we will write  $\mathbb{B}^{(k)}$ . One solution would be simply take the  $L$  largest values; alterna-

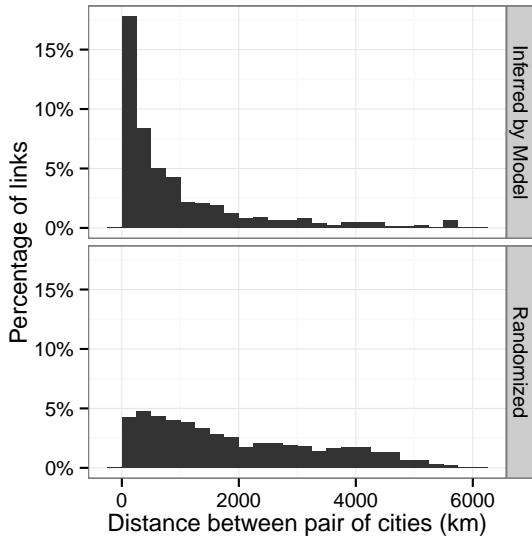


Figure 5.6: Histograms of distances between pairs of connected cities, in model-inferred networks (top), versus “negative” networks from  $Q$  (bottom).

tively, we could take the  $L$  coefficients for which we are most confident that  $a_{r_1, r_2}^{(k)} > 0$ . We strike a balance between these two extremes by sorting the dynamics coefficients according to the lower bound of their 95% confidence intervals. This ensures that we get city pairs for which  $a_{r_1, r_2}^{(k)}$  is significantly distinct from zero, but that we also emphasize large values rather than small values with low variance. Per-sample confidence intervals are obtained by computing the closed form solution to the posterior distribution over each dynamics coefficient,  $P(a_{r_1, r_2}^{(k)} | \eta_{r_1}^{(k)}, \eta_{r_2}^{(k)}, \lambda)$ , which, in ridge regression, is normally distributed. We can then compute the 95% confidence interval of the coefficients in each  $A^{(k)}$ , and sort them by the bottom of this confidence interval,  $\tilde{a}_{i,j}^{(k)} = \mu_{a_{i,j}^{(k)}} - Z_{(.975)}\sigma_{a_{i,j}^{(k)}}^2$ , where  $Z_{(.975)}$  is the inverse Normal cumulative density function evaluated at 0.975,  $Z_{(.975)} = 1.96$ . We select  $L$  by the number of coefficients that pass the  $p < 0.05$  false discovery rate threshold in the aggregated network ( $L = 510$ ), as described in the previous section. This procedure yields  $K = 100$  different discretized influence networks  $\bar{\mathbb{B}}^{(k)}$ , each with identical density to the aggregated network  $\bar{\mathbb{B}}$ . By comparing the logistic regression coefficients obtained within each of these  $K$  networks, it is possible to quantify the effect of uncertainty about  $\eta$  on the substantive inferences that we would like to draw about the diffusion of language change.

## 5.4 Results

Figure 5.7 shows the resulting logistic regression coefficients. While geographical distance is prominent, the absolute difference in the proportion of African Americans is the strongest predictor: the more similar two metropolitan areas are in terms of this demographic, the more likely that linguistic influence is transmitted between them. Absolute difference in the proportion of Hispanics, residents of urbanized areas, and median income are also strong predictors. This indicates that while language change does spread geographically, demographics play a central role, and nearby cities may remain linguistically distinct if they differ demographically, particularly in

	linked mean	linked s.e.	nonlinked mean	nonlinked s.e.
<i>geography</i>				
distance (km)	919	36.5	1940	28.6
<i>symmetric</i>				
abs diff % urbanized	9.09	0.246	13.2	0.215
abs diff log median income	0.163	0.00421	0.224	0.00356
abs diff median age	2.79	0.0790	3.54	0.0763
abs diff % renter	4.72	0.132	5.38	0.103
abs diff % af. am	6.19	0.175	14.7	0.232
abs diff % hispanic	10.1	0.375	20.2	0.530
<i>asymmetric</i>				
raw diff log population	0.247	0.0246	-0.0127	0.00961
raw diff % urbanized	1.77	0.389	-0.0912	0.112
raw diff log median income	0.0320	0.00654	-0.00166	0.00187
raw diff median age	-0.198	0.113	-0.00449	0.0296
raw diff % renter	0.316	0.195	-0.00239	0.0473
raw diff % af. am	0.00292	0.244	0.00712	0.109
raw diff % hispanic	0.0327	0.472	0.0274	0.182

Table 5.3: Differences between linked and (sampled) non-linked pairs of cities, summarized by their mean and its standard error.

terms of race. African American English differs more substantially from other American varieties than any regional dialect (Wolfram and Schilling-Estes, 2005); our analysis suggests that such differences persist in the virtual and disembodied realm of social media. Examples of linguistically linked city pairs that are distant but demographically similar include Washington D.C. and New Orleans (high proportions of African-Americans), Los Angeles and Miami (high proportions of Hispanics), and Boston and Seattle (relatively few minorities, compared with other large cities).

Of the asymmetric features, population is the most informative, as larger cities are more likely to transmit to smaller ones. In the induced network of linguistic influence  $\mathbb{B}_{0.05}$ , the three largest metropolitan areas – New York, Los Angeles, and Chicago – have 40 outgoing connections and only fifteen incoming connections. Wealthier and younger cities are also significantly more likely to lead than to follow. While this may seem to conflict with earlier findings that language change often originates from the working class, wealthy *cities* must be differentiated from wealthy *individuals*: wealthy cities may indeed be the home to the upwardly-mobile working class that Labov associates with linguistic creativity (Labov, 2001), even if they also host a greater-than-average number of very wealthy individuals.

Additional validation for the logistic regression is obtained by measuring its cross-validated predictive accuracy. For each of the  $K$  samples, we randomly select 10% of the instances (positive or negative city pairs) as a held-out test set, and fit the logistic regression on the other 90%. For each city pair in the test set, the logistic regression predicts whether a link exists, and we check the prediction against whether the directed pair is present in  $\mathbb{B}^{(k)}$ . Results are shown in Table 5.4. Since the number of positive and negative instances are equal, a random baseline would achieve 50% accuracy. A classifier that uses only geography and population (the two components of the gravity model) gives 66.5% predictive accuracy. The addition of demographic features (both asymmetric and symmetric) increases this substantially, to 74.4%. While symmetric features obtain the most

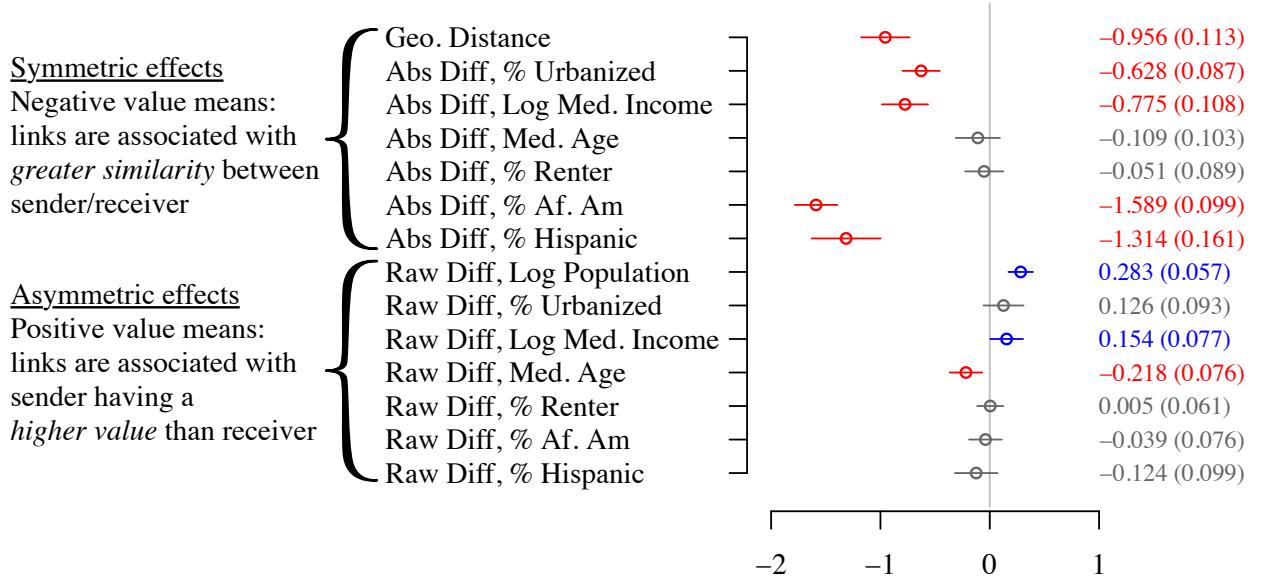


Figure 5.7: Logistic regression coefficients for predicting links between city (MSA) pairs. 95% confidence intervals are plotted; standard errors are in parentheses. Coefficient values are from standardized inputs; the mean and standard deviations are shown to the right.

robust regression coefficients, adding the asymmetric features increases the predictive accuracy from 74.1% to 74.4%, a small but statistically significant difference.

	mean acc	std. err
geography + symmetric + asymmetric	74.37	0.08
geography + symmetric	74.09	0.07
symmetric + asymmetric	73.13	0.08
geography + population	67.33	0.08
geography	66.48	0.09

Table 5.4: Average accuracy predicting links between MSA pairs, and its Monte Carlo standard error (calculated from  $K = 100$  simulation samples). The feature groups are defined in Table 5.3; “population” refers to “raw diff log population.”

## 5.5 Discussion

Language continues to evolve in social media. By tracking the popularity of words over time and space, we can harness large-scale data to uncover the hidden structure of language change. We find a remarkably strong role for demographics, particularly as our analysis is centered on a geographical grouping of individual users. Language change is significantly more likely to be transmitted between demographically-similar areas, especially with regard to race — although

demographic properties such as socioeconomic class may be more difficult to assess from census statistics.

Language change spreads across social network connections, and it is well known that the social networks that matter for language change are often strongly homophilous in terms of both demographics and geography (Milroy, 1991; Labov, 2001). This chapter approaches homophily from a macro-level perspective: rather than homophily between individual speakers (Kwak et al., 2010), we identify homophily between geographical communities as an important factor driving the observable diffusion of lexical change. Individuals who are geographically proximate will indeed be more likely to share social network connections (Sadilek et al., 2012), so the role of geography in our analysis is not difficult to explain. But more surprising is the role of demographics, since it is unclear whether individuals who live in cities that are geographically distant but demographically similar will be likely to share a social network connection. Previous work has shown that friendship links on Facebook are racially homophilous (Chang et al., 2010), but to our knowledge the interaction with geography has not been explored. In principle, a large-scale analysis of social network links on Twitter or some other platform could shed light on this question. Such sites impose restrictions that make social networks difficult to acquire, but one possible approach would be to try to link the “reply trees” considered by Gonçalves et al. (2011) with the geographic and demographic metadata considered here; while intriguing, this is outside the scope of the present chapter. A major methodological contribution is that similar macro-scale social phenomena can be inferred directly from spatiotemporal word counts, even without access to individual social networks.

Our approach can be refined in several ways. We gain robustness by choosing metropolitan areas as the basic units of analysis, but measuring word frequencies among sub-communities or individuals could shed light on linguistic diversity *within* metropolitan areas. Similarly, estimation is facilitated by fitting a single first-order dynamics matrix across all words, but some regions may exert more or less influence for different types of words, and a more flexible model of temporal dynamics might yield additional insights. Finally, language change occurs at many different levels, ranging from orthography to syntax and pragmatics. This work pertains only to word frequencies, but future work might consider structural changes, such as the phonetical process resulting in the transcription of *i don't* into *ion*.

It is inevitable that the norms of written language must change to accommodate the new ways in which writing is used. As with all language changes, innovation must be transmitted between real language users, ultimately grounding out in countless individual decisions — conscious or not — about whether to use a new linguistic form. Traditional sociolinguistics has produced many insights from the close analysis of a relatively small number of variables. Analysis of large-scale social media data offers a new, complementary methodology by aggregating the linguistic decisions of millions of individuals.

## 5.6 Appendix: Data processing

We perform several preprocessing steps to prepare the raw Twitter feed for analysis, described in this section: (1) a message preprocessing pipeline, and (2) a selection procedure for the words to analyze. Supplementary files are available at:

<http://brenocon.com/DiffusionOfLexicalChangeInSocialMedia>

Filenames in this section refer to paths within this resource.

### 5.6.1 Messages

Our initial dataset is of Twitter Gardenhose/Decahose messages from August 2009 through September 2012, containing approximately 17 billion tweets. 721 million were found to have a geotag, and 171 million were located in the United States. After MSA and content filtering, 107 million messages (from 2.7 million unique user accounts) remained for the analysis. The preprocessing software is available at [https://github.com/brendano/twitter\\_geo\\_prepoc/](https://github.com/brendano/twitter_geo_prepoc/) and a copy is archived as supplementary information file [preprocessing\\_pipeline.zip](#).

**Geotags** The Twitter API’s structured data includes a field for latitude and longitude coordinates from users who have enabled geo-location; typically, these come from messages authored on mobile phones. Besides that field, there are also informal geotags in the *user.location* field, from clients that insert coordinates as a string; for example, *ÜT: 40.043883,-88.275849* is a geotag from the ÜberTwitter client. These informal geotags are more common in earlier data, and are the only source of coordinates before Twitter added official support for coordinate geotags in late 2009. A regular expression extracts this type of coordinates; there were about twice as many messages with informal coordinates as messages with official API coordinates. We use both types of messages.

**Location** We use only messages from the continental USA, locating the latitude and longitude coordinates to a county or county-equivalent, according to the U.S. Census Bureau’s 2010 TIGER/Line Shapefiles. (<http://www.census.gov/geo/maps-data/data/tiger-line.html>). The United States Office of Management and Budget defines a set of Metropolitan Statistical Areas (MSAs), which are not legal administrative divisions, but rather, geographical regions centered around a single urban core (Office of Management and Budget (USA), 2010); every MSA is defined as a set of counties. We consider the 200 most populous MSAs in the lower 48 U.S. states. The most populous MSA is centered on New York City (population 19 million); the 200th most populous is Fargo, North Dakota (population 200,000). We retain messages whose location belongs to one of these MSAs. According to the 2010 census, the 200 largest MSAs include 76% of all US residents in the lower 48 states; however, we find that these MSAs contain 89% of all Twitter messages sent from within the lower 48 states, which coheres with recent work showing that geotagged Tweets are more likely to come from urban areas (Hecht and Stephens, 2014).

For each MSA, demographic attributes are computed from the 2010 U.S. Census. The following demographic attributes are included: log population, log median income, % residents in urbanized areas, media age, % renters, % African American; % Hispanic. We did not consider % European American because it has a strong negative correlation with % African American,  $r = -0.71$ ; we did not consider % Asian American because it is much smaller, with a median value of 2.8%. Mean and standard deviations of all demographic attributes are shown in Table 2 of the main text.

**Content and Follower Filtering** Several additional processing steps were then performed to remove marketing-oriented and spam accounts. We remove all messages written by users who have more than 1000 followers, or who follow more than 1000 people. This helps to eliminate automated accounts, particularly content polluters (Lee et al., 2011). We remove all messages that are retweets—either marked as such in the API’s structured data, or any message containing the word *RT* (in either lowercase or uppercase). While retweeting could be a useful linguistic signal in its own right, we prefer to focus on original text. Finally, any message containing a URL is removed; this acts as a filter to remove automated and marketing-oriented content, which is typically designed to draw the reader to a page elsewhere on the web. Of course, these filters also

eliminate some legitimate messages, but since there is no shortage of data, we prefer to focus on a subsample that is more likely to contain original, non-automated content.

**Time** Each Twitter message includes a timestamp. We aggregate messages into seven-day intervals, which facilitates computation and removes any day-of-week effects. Each interval starts on Monday at UTC 0800, corresponding to 12am PST and 4am EDT.

## 5.6.2 Words

To select the set of words to analyze, we begin with the 100,000 most frequent terms, excluding hashtags and usernames. We further require that each term must be used more than more than twenty times in ten different metropolitan areas. We compute the variance of the word's log probability over time ( $\nu_{w,t}$  in Equation 2 of the main text, estimated in a standalone step, as described there), and require that the variance be greater than three. This cutoff was chosen so that roughly 5,000 words would be selected; we end up with 4,854 words. From this subset, we manually eliminate all named entities and non-English words. This determination is ambiguous because some strings can reference both names and words (e.g. *homer*, a dictionary word that often references the character *Homer Simpson*) or multiple languages (e.g. *y*, which can mean *and* in Spanish, and *why* in informal English). For each term, we randomly select twenty example messages and manually determine from context whether the usage is as an English word. We retain terms that are used as English non-name words in at least 80% of the examples.

The final word set contains 2,603 words. Our annotation decisions for all 4,854 words can be seen in our supplementary information file, [name\\_annotations.tsv](#), and the selected words can be seen in Figure 5.8 (or the file [wordlist\\_table.pdf](#)). The usage examples we inspected are available in [word\\_examples\\_for\\_annotation\\_in\\_cluster\\_order.html](#).

The overall results of our analysis are broadly similar when we do not perform manual word filtering, but this filtering enables us to focus on changes in (English) language rather than in the popularity of entities or in the overall multilingual composition of American Twitter users.

All text was tokenized using the *Twokenize.java* program, which can be downloaded at <http://www.ark.cs.cmu.edu/TweetNLP/>. *Twokenize* is designed to be robust to social media phenomena that confuse other tokenizers, such as emoticons (O'Connor et al., 2010c; Owoputi et al., 2013). Repetition of the same character two or more times was normalized to just two (e.g. *sooooo* → *soo*). No other preprocessing (e.g., stemming) was performed.

hheeyy	u'd	warms	twerk	qo	it-	leavn
niccee	iwanna	heats	tango	kum	thiiis	droppin
oommgg	imaa	wlk	dougie	kome	dhis	chargin
yyaayy	u'll	plow	pik	wait	thys	skippin
bomb.com	culd	re-up	mute	elses	dhat	leavin
=d	cld	swerve	twit	happn	thiz	guarding
=p	cud	beez	twitt	i`m	whr	stealin
=o	shud	hangout	twitvid	iim	wut	sendin
gooaall	shuld	livee	shovel	u're	waht	pushin
gooll	shld	workin	fite	something's	wher	usin
shee	mite	wrking	knit	iits	whut	catchin
uve	wud	wrkn	dribble	itz	weneva	breakin
u've	wuld	wrkin	fling	nothing's	wha	bringin
jut	wld	workn	subtweet	everything's	yy	cuttin
jux	iont	sittin	fuxx	thatz	veryy	pickin
jus	dn't	standin	hitt	datz	absolutly	carryin
jsut	wldnt	stayn	bodied	thas	completly	clockin
iaint	wulndt	stayin	toot	dass	awkwardly	keepn
ain't	shudnt	layin	mow	whas	legitimately	holdin
aint	wudnt	sittn	go2	watz	uber	throwin
ainn	shldnt	chattin	kall	wuts	f*ckin	feedin
aiint	cudnt	dealin	sext	wutz	fuggin	showin
couldve	kant	steppin	textt	whts	f-ing	puttn
wulda	culdnt	otp	fugg	wats	fukkin	hittin
shulda	cldnt	relaxin	gtf	whatss	fuccin	shuttin
wouldve	hafta	checkn	fcuk	whatz	effn	addin
shouldve	gotta	cashin	wink	wutchu	mf'n	scratchin
wudda	knoo	cuddling	smooches	watchu	flippin	switchin
shudda	noe	checkin	shrugs	nobody's	f'n	choppin
wuda	warmers	flirtin	sniffles	smellin	f'in	tearin
shuda	excite	grillin	faints	feeln	effin	givin
mustve	beleive	mackin	kanyeshrug	feelin	s0	rubbini
nevr	gaf	arguin	squee	soundin	liike	puttin
neva	4get	mobbin	pouts	dressin	liek	settin
eva	rememba	chillaxin	sideeye	wus	lk	lickin
evr	memba	wakin	winks	waz	boutt	postin
evar	luvv	gearing	shruggs	wass	havin	hittin
onli	luv	waken	grins	wuz	havn	touchin
evn	nominat	snuggled	shrug	iz	beinq	tackling
eem	h8	turnt	ahem	izz	get'n	installing
realli	thnk	fukked	mjb	:'s	gettn	takin
alrdy	thght	racked	sadface	buhh	follown	expectin
cuda	4got	chk	pause	pero	followin	wearin
allready	thk	checc	cosign	esp	callin	takn
alreadyy	quess	shouts	syh	i.e.	treatin	despicable
offically	gotchuu	matata	fwm	i.e.	helpin	buyin
finaly	swea	d/l	holla	iwant	lettn	makin
ion	hav	preorder	hmu	gimmie	gimn	makn
uma	qot	occupy	co-sign	whys	judgin	watchinq
casually	favorited	pre-order	laff	imiss	lettin	watch'n
deadass	hadd	nano	likey	ilovee	unfollowin	watchn
obvi	pre-ordered	sync	muero	sumtimes	testin	watchin
alwayz	copped	conditioned	clinch	evrytime	killn	hearin
alwys	preordere	shun	reinstall	iguess	answerin	findin
definitly	aced	twug	install	mayb	teachin	likin
defiantly	missd	slander	decorate	altho	stalkin	surfin
definetly	askd	scramble	transform	methinks	calln	blastin
definately	tol	punt	tackle	becuse	askn	luvin
gonna	bbm'd	flee	discover	b/c	askin	enjoyin
bouta	calld	draw	violate	kause	bbm'n	readin
fina	startd	sweep	deactivate	wheneva	teln	downloadin
gne	sed	spill	followback	becuz	stoppin	rockn
trynaa	knos	sabotage	qet	eventhough	losin	celebratin
fenna	knws	reboot	takee	what're	changin	draggin
letss	luvs	wobble	makee	iif	updating	grabbin
letz	forgives	swim	giv		payin	coppin

Figure 5.8: (Page 1 of 6): All 2,603 words used in our main analysis. They are ordered by the hierarchical word clusters of (Owoputi et al., 2013) (<http://www.ark.cs.cmu.edu/TweetNLP/>) which tends to group words with similar syntactic or semantic properties. The lowercased forms are shown, which sometimes is not the most common form; for example, “:D” is usually written as “D”.

smashin	some1	swagged	bak	dancin	editing	again
duin	evryone	dunked	baik	ringin	blogging	18r
doinn	evry1	lookd	baq	sleepn	unpacking	lata
doiin	every1	swam	bacc	beastin	syncing	sumtime
doin	evrybdy	sacked	bac	snitchin	scanning	nomo
doinq	everyonee	bullied	right	cryin	grading	2u
eatin	evrybody	benched	righ	hollin	decorating	tbh
grilling	oomf	nominated	riite	flopping	knitting	ftl
cookin	oomfs	ranked	rite	breathin	coding	afterall
eattin	meeka	snowed	rght	twerkin	designing	mehn
bakin	no1	installed	ritee	choosin	pitching	nshit
cookin	whoeva	launched	rii	laffin	floodng	jor
drinkn	waitn	unplugged	ryte	twerking	marvins	neways
cravin	waitin	leaked	schemin	laughn	marvin's	leh
drinkin	rootin	hosted	hydrated	laughin	openin	anywayz
mert	feenin	stung	storming	stylin	mowing	nemore
mehh	searchin	playd	rainin	swaggin	shoveling	neway
mhe	lookn	hoed	poppin	trickin	cleanin	2me
urself	look'n	subtweeted	snowing	performin	stocking	lah
yaself	commuin	sampled	poppington	pumpin	cooling	4me
yurself	comin	wantd	happennin	flexin	finishin	2m
hym	comm	4ward	goodie	starin	signin	yest
ypu	cummin	starvin	poppn	hooping	fillin	4u
yall	upgrading	singlee	snownin	spinnin	wrapping	lastnite
y'all	startn	outtie	popin	blazin	washin	evryday
yeen	startin	pregggers	hailing	jerkin	cashing	errday
iget	plannin	siick	wrong	weetin	passin	rn
yhuu	omw	sauced	premieres	planking	peeling	tmo
yhu	headin	faded	rox	cuffin	drivin	yesturday
y0u	enroute	preg	suxx	packin	travelling	doee
juu	s'\o	sunburned	sux	studyin	dunking	thoe
foy	s/o	sunburnt	snows	trolling	walkn	doe
ilt	shoutouts	tite	wrks	writin	beaming	thoo
iit	slo	odee	tackles	subtweeting	walkin	thoee
yurs	s/0	maad	scrolls	twittering	shuffling	thoughh
urs	qoinq	od	presents	protesting	up-	now-
everything	fixin	madd	mower	grindin	uhp	ther
everythin	crackn	embarassed	toasty	bbming	uprt	heaa
everythin	jumpin	appalled	cozy	chokin	owt	4eva
nuttin	shakin	dissapointed	warm	fightin	Out	4ever
nuttin	rushin	surprised	rigged	texting	out-	sumwhere
nuttin	crackin	butthurt	postponed	fasting	ova	forsure
nuffin	goin	dgaf	trendin	twitting	ovr	manually
nuttin	movin	xcited	undefeated	recordin	ovaa	4sure
nutin	qoin	talm	deadd	graduating	ovah	lyrically
nuthin	stickin	talk'n	saucer	sexin	arnd	eitha
nothn	listenin	tlkn	2go	spellin	outsidee	.
nuthin	willin	bitching	bk	studying	w/u	:
nuthn	sposed	talkn	tangled	blockin	w/me	;
nthn	refering	braggin	doobies	singin	2gether	?
nun	cooled	talkinq	sweatin	mixin	2getha	?..
sumtin	mowed	complainin	slackin	sharin	2morro	?
sumthn	deactivated	speakin	chirping	slammin	2morrow	??
sumthing	spendin	thinkn	thuggin	twatchin	tomar	?!
somethn	wastin	thnkn	partyin	datin	2mrw	!!
smthn	violated	talmbout	lien	cuffing	2mor	!
summin	dvr'd	jokin	wildin	typin	2mrrw	..
sumthin	muted	forgettin	knockin	smackin	2moro	!!.
suttin	graduated	claimin	buzzin	subtweetin	2maro	ndd
anythin	bugged	dunno	cheatin	scheming	toma	&
nethng	swept	wonderin	frontin	hidin	2morow	&&
nebody	walkd	debatin	buggin	flexing	2day	(&
any1	rained	pondering	dyin	twatching	2nite	be4
ne1	fouled	guessin	speedin	learnin	tonite	b4
sumone	biked	hopin	preachin	uploading	tnite	witout
sum1	ducked	knowin	coughin	writting	2night	afr

Figure 5.8. Page 2 of 6.

afta	mem	heehee	whelp	imyy	8-)	:/
t0	supp	jaja	hmpf	toodles	>d	:*(
of	vox	ahaa	hmp	imu	;p	=["
4the	nit	jajaja	uugh	hbd	>)	:-"
w/a	ren	hihi	uqhh	plzz	:3	:("
w/my	ff	jajajajajaja	urghh	hunh	(;	-"
w/the	det	wkwkwk	urgh	abi	toort	><
w/this	sk	jajajaj	hmmpf	hbu	^.^	:((
widd	aff	jaja]	welp	wbu	=)	>--<
w	lat	inshallah	uuggh	wby	=])	""
unda	stu	yuup	uqh	whassup	:}	lok
wiit	rm	yehh	grr	wassupp	(:	hakuna
frum	wr	nawl	wowzers	wussup	(:	love-
frm	hu	noes	arrgh	goodlookin	^_^	(*
4rm	int	werd	thanxx	wydd	=])	*)
n2	rb	iknow	thankx	tf	:")	chuuch
in2	ent	nawh	thnx	wusup	^_^	gnr
btwn	thr	yh	thankz	df	((;	foh
durin	u's	wordd	thanx	wzup	^_^-^	smhh
w/in	gz	ikr	waddup	((:	tyna	
On	sp	saludos	wadup	;)()	smfh	
iin	baken	ayye	thansx	wattup	;)()	fyl
i'n	ft	yuupp	thnxx	4real	cx	kmt
tge	fah	asdfghjkl	srry	wuddup	:-.	i/k
tha	fir	omo	sry	watup	haga	guhh
dha	ov	ooww	felicidades	whatup	o__o	rns
onna	w/that	waahh	grats	wydt	u_u	goodtimes
inda	w/her	wheww	congratz	wya	o__o	frfr
2my	mos	oww	twugs	whatup	0_0	iswear
yhur	hurd	owee	booyah	wsup	:o	ijjs
urr	derp	wooww	hooray	ik	o_0	j/p
beyonces	lml	zomg	whoohooo	iono	0__o	(:<
yurr	lbvs	ayee	yummmyy	idek	o.o	<<"
ure	lolx	uugggh	yipee	iknw	(---)	<
nicki's	lolss	wheeww	woo-hoo	becareful	o_o	<-
somebodys	lls	ggrr	hoo	idgaf	0_o	:
google's	lol	omq	whee	amo	x_x	
rihanna's	lolzz	p.s.	leggo	amoo	o:	<~
beyonce's	lols	awee	ftw	whatev	:oo	<==
hiz	wkwk	awhh	huzzah	ig	:/	'
year's	me2	imy2	woot	wuteva	:"(	<<
blog	rft	oww	yuumm	busta	</3	lol-
headlines	rotflmao	oke	muah	whatev	):	\3
downloads	yessirr	aaww	tgif	gtfo	:-(	^_-
chronicle	rofl	awh	w00t	aight	:-	3
q&a	kml	wahh	gobble	aite	:-.	^
index	sheeshh	tuhh	rah	ard	:-/	np-
cc	ctfu	awl	burr	iite	:(	drake-
fw	ctfuu	uumm	meow	ookk	=	viendo
sn	leggoo	huuhh	gudda	s2	:-"	>
psa	roflmao	hhmm	ayo	<33333333	=	voice)
attn	legoo	kthx	cmon	<333	--	}
sidenote	rotfl	sheesh	brr	<3333333	:-\	>>"
np	kmsl	brb	gmorning	<3<3<3	>.<	>
wts	bol	whew	g'morning	<33	=/	::
null	sheesshh	nbd	laters	<33333	:-"	>>
cont	hyfr	arghh	g'nite	<3<3	=	merry
via	yezzir	argh	gnight	<33333	:-	jagged
itrt	4sho	geezz	flyy	<3333	-	/mi
urt	lma0	uughh	ilh	b-)	:/	sportacular
fb	rotf	wheew	gnite	;3	>:o	]):
swine	bwahahahaha	blech	iloveyou	*.*	,-	[
*	jajajajaja	geesh	imy	u.u	,-	f/
punya	jajajajaja	ick	goodnite	\m/	x_x	yg
cade	aha	arrgghh	g'night	;	:/	ng
pow	tqm	arg	goodmorningg	tttyl	<,_<	

Figure 5.8. Page 3 of 6.

/via	humps	timeline	tweoples	cookouts	foams	night
/cc	boobie	cuzzins	twits	alphas	nudes	moorning
hahart	tweep	homegirl	tweeple	spammers	subtweets	mornin
smhrt	jeezy	tl	twitterverse	spiders	timelines	morntng
lolrt	9700	boothang	twam	macs	mentions	mawnin
lmaort	heffa	homegurl	twiggaz	savages	twitpics	morningq
oan	nigha	bestfriendd	tworld	tornados	e-mails	morn
inches	bih	besty	ya'll!	bigs	twitcons	day-
chainz	nikka	twifey	bloggers	seniors	avi's	semester
wks	nicca	meech	troops	vampires	skillz	wk
loko	nukka	roomate	commentators	turntables	viruses	decade
lokos	nucca	co-worker	grads	fakers	piles	wknd
tds	bihh	followerss	promoters	referees	pix	shidd
rebounds	twiggga	follwers	miners	allergies	wrds	shytt
assists	wuss	fren	interns	jumpers	vowels	schoolwork
innings	g6	bff's	graduates	blisters	disappointme	taxes
td's	spammer	bestfriends	travelers	brackets	nts	hw
pts	slacker	thngs	execs	sacks	turnovers	homework
yards	bish	ladys	crews	snapbacks	fouls	hmwk
flags	mfer	lass	riots	bikinis	leagues	errands
yds	groupie	girlz	dj's	weaves	tryouts	enuf
km	bumb	nupes	bots	dunks	meetings	enuff
min/mile	cornball	gurlz	unions	helmets	medals	hickies
ln	nupe	ratchets	shoppers	scarves	pools	matta
utc	heaux	refs	users	mints	camps	tyme
pst	dubb	shorties	tablets	beatz	penalties	tiime
hunnit	thugg	gurlz	djs	quizzes	pitches	thng
veces	dweeb	bruhs	producers	exams	finals	thing
yd	hoee	sistas	qbs	wigs	touchdowns	werk
cpl	mf	grls	mc's	toys	vids	dreamland
baybee	qirl	niccias	stans	stockings	billboards	k.o
hommie	gyal	mf's	candidates	sweatpants	lists	multitask
hunnie	grl	suckas	developers	textbooks	cribs	intermission
hunni	dag	heffas	receivers	hoodies	badges	midnite
hunn	wrd	heauxs	accts	boots	dvd's	soundcheck
sandz	guh	shyts	haitians	sweaters	files	fantasy
homey	kellz	mfs	republicans	coats	avatars	halftime
booskie	shid	mofo's	hackers	cd's	hashtags	h.s.
brahh	shiid	hoes	libras	uniforms	blogs	tim
hun	bwoy	bishes	leos	mixtapes	ringtones	regionals
siss	manee	nighas	protesters	dollaz	functionality	skewl
dawgg	b-day	nikkas	jamaicans	lacefronts	mixes	skoo
booski	bdayy	chics	scorpios	fitteds	beaches	skool
mamacita	gabba	guyz	celebs	pins	presentations	twitterjail
sweety	gambino	twiftfam	virgos	pumps	debates	recital
famo	axx	pplz	bikers	dreds	podcasts	class
sonn	asx	guise	mosquitos	cargos	icons	exam
hon	arenas	twitterworld	crooks	tees	polls	orientation
soror	daddys	twitters	tornadoes	blankets	dvds	midterm
baee	neos	twitches	roads	tights	forums	homeroom
meng	roomates	twitts	lasers	sandals	quo	klass
bestiee	homegirls	ya'll!!	fireworks	essays	feedback	reherusal
cuhh	exs	twopeople	barbies	boyshorts	updates	tamales
lovie	eyez	girlies	mosquitoes	joints	tix	gifts
babez	sideburns	twamly	turkeys	hitters	tixs	wedges
bae	jammies	tweeple	tt's	costumes	invites	gumbo
bruhh	talents	twiggas	tts	emojis	suggestions	crawfish
bby	grades	tweeps	c's	leggings	remedies	blueberries
ddub	cockiness	peeps	midterms	laptops	takers	smoothies
bbs	butthole	tweetharts	sigmas	gloves	predictions	pumpkin
tete	momz	twitterville	professors	cleats	nites	chowder
bb	cuzzin	weeties	earthquakes	trunks	night	cherries
tunechi	novio	lovies	kiddies	grenades	nyte	casserole
todos	bestfriend	twitterland	deltas	lighters	nite	ice-cream
babii	homeboi	hunnies	bats	trax	nitee	barbecue
lul	cuzzos	sorors	subliminals	snuggies		sandwhich

Figure 5.8. Page 4 of 6.

canes	warmth	sledding	mamba	concussion	banga	documentati
bbq	capitalism	swimming	sleigh	smut	tornado	n
locos	ambition	golfing	crackberry	tux	shutdown	conversion
turkey	ridiculousness	h.a.m	comp	badger	fumble	exp
veg	sucess	snowboarding	trampoline	peacock	grenade	integration
chilli	snow	biking	vm	slime	layup	ui
watermelon	pollen	shoppin	router	broom	lapdance	interface
ceasar	sleet	skiing	trackball	bootz	touchdown	auditorium
choc	flurries	tanning	bberry	reindeer	homerun	bldg
frap	thunder	bowlin	fone	firework	shoewrr	blitz
frappe	lightening	fishin	fne	tings	spliff	conditioning
butta	thunderstorm	swimmin	fones	tt	salts	heater
smores	s	clubbin	background	snowball	texter	fest
popsicles	storms	kayaking	avi	snowman	snippet	festival
potatoe	willies	camping	icon	cooter	screenshot	klub
pina	fog	tubing	avii	mosquito	vid	medal
chix	leopard	kickback	twiticon	beeper	hitter	computing
pneumonia	showers	potluck	avatar	fitted	3some	disasters
bronchitis	perf	cookout	background	hipster	mixtape	bases
libra	ratchetness	bane	acct	monsta	cypher	scorer
terrorism	goody	begining	acc	rockstar	joint	beatdown
bullying	coonery	weeknd	layout	jumpoff	sextape	afterparty
tsunami	fiyah	wrld	default	ting	song	presentation
qualifying	shiznit	abyss	twitcon	b+	bulpen	banquet
3-d	bitchassness	motto	gamertag	lacefront	playoffs	sesh
detention	bullshyt	creator	namee	bikini	tournament	gala
3d	piff	miz	acct	mink	matchup	conf
hd	lrt	crazies	acc	guido	lockout	keynote
trainin	ping	buzzer	tethering	d*ck	redzone	tweetup
threes	autocomplete	sideline	firmware	choppa	balggame	photoshoot
persuasion	twitted	ballpark	widget	professor	topic	webinar
nightlife	twitterr	caf	api	ump	af	ceremony
defense	deck	hospt	browsr	prof	texters	finale
amusement	campus	sidelines	site	pin	a'f	demo
comm	racks	dancefloor	homepage	fams	ASF	premiere
psychology	hoarders	beachh	plugin	bottoms	mgr	catalog
accounting	rss	moviess	earthquake	ribbon	coordinator	opener
anatomy	dsl	bookstore	assassin	lite	commerical	photog
mojito	a/c	supermarket	umpire	coat	pact	kicker
chardonnay	apocalypse	pool	outburst	hoody	forum	goalie
merlot	physics	bachelor	essay	trench	prototype	quarterback
moscato	calc	endzone	e-mail	cardigan	sidebar	bracket
eggnog	astronomy	studio	spam	snapback	nomination	receiver
joose	sociology	dugout	interception	jacket	cache	o-line
martini	calculus	carwash	emoji	sweatshirt	flyer	pitcher
cider	omega	in-laws	allergy	gown	software	officiating
patron	geometry	fireplace	alchy	stylist	wiki	qb
liquor	algebra	fireplace	icee	sundress	scandal	dictator
mojitos	math	internets	wave	scarf	portfolio	promoter
latte	pc	library	muzik	hoodie	database	hoax
nuvo	psych	patio	musik	v-neck	newsletter	client
lipgloss	thesis	syllabus	bootleg	sweater	interference	vaccine
guestlist	chem	buildin	disk	hammock	savings	producer
sunscreen	quince	it is	webcam	mound	decor	refill
lint	biology	bizness	batt	parade	elections	study
radiation	creole	grinch	antenna	maze	svc	setup
bandwidth	lte	wackness	lab	licker	processor	plank
drilling	tablet	juiceman	modem	cabin	panels	file
tide	froyo	porch	cpu	tree	outage	rebound
h20	jailbroken	sunroof	netbook	recession	rankings	dunk
randomness	desktop	bizz	adapter	volcano	themes	lecture
autotune	gametime	lawn	gift	storm	supplies	scrimmage
climax	bfast	crib	promo	breeze	workflow	turnover
sunburn	poolside	babyshower	swagga	restaurant	developer	protest
applause	dorm	dorm	wasp	swimsuit	tutoring	check-in
thirst	stepfather	stepfather	quickie	thunderstorm	revolt	

Figure 5.8. Page 5 of 6.

tailgate	low-key	foggy	rnb	nigthy
rally	quik	humid	kick-ass	nitey
timeout	krazy	snowy	burlesque	lst
shootout	wackk	coldd	horror	nxt
wrkout	cray	muggy	snowboard	nex
kickoff	crazi	chilly	glam	farmer's
tutor	crazi	thundering	camp	awkward
retreat	crazzy	coold	dubstep	neww
meanin	crzy	freezin	sci-fi	wittle
scope	smoove	tamed	hiphop	biig
launch	nutz	swizz	lotus	biq
bundle	fancy	linen	harvest	teeny
disclaimer	wack	flannel	mvp	24hr
s/n	trill	camo	hip-hop	womans
verdict	trippy	suede	westcoast	meteor
ques	stoopid	pajama	soca	uur
url	wreckless	fleece	vamp	otha
vow	wak	jingle	fg	othr
hashtag	grimey	swamp	bronze	2c
tribute	judgmental	gingerbread	shark	t2
refund	rediculous	blk	lightning	tew
meetin	annoyn	blk	lib	2b
mtg	borin	haunted	football	2da
qtr	awk	geaux	pre-season	2the
inning	embarassing	twitterless	crossfit	2do
quater	spooky	balloon	b-ball	2get
warranty	hopeless	atta	triathlon	ths
resolution	debatable	olympic	allstar	yesterdays
costume	irrelevant	women's	frisbee	anotha
roulette	halala	iconic	spades	anutha
decorations	hillarious	unsung	gameday	evry
resolutions	unforgettable	womens	lacrosse	a\$ap
gully	amazin	mens	multiplayer	n0
dutty	halarious	upscale	preseason	
wavy	truue	celeb	soccer	
trendy	truu	freelance	baseball	
matchin	nicee	prepaid	varsity	
ratchet	kewl	wireless	golf	
sexii	niccee	touchscreen	pep	
rachet	hott	seasonal	fball	
dtf	kute	mindless	vball	
beastly	gudd	xtra	f1	
funnel	gd	external	softball	
peppermint	qood	batting	derby	
corned	guud	static	playoff	
tanned	gr8	curved	volleyball	
bizzy	grreat	mp3	ski	
chapped	kickass	torrent	kickball	
collard	niice	pdf	postseason	
clutch	doobie	rbi	stimulus	
intresting	precious	offical	healthcare	
interestin	2much	infamous	hc	
ode	betta	domain	voter	
ez	bettr	dr's	immigration	
prosperous	bettah	parkin	reform	
diff	colder	secondary	web	
dif	warmer	grad	cust	
seperate	hotter	flu	biz	
ppl's	sharper	quake	gadget	
subliminal	long	couture	baskets	
ppls	tardy	farmers	skating	
winky	windy	rack	quiz	
foul	dreary	powerhouse	overcast	
futuristic	cold	renegade	wackest	
shameless	rainy	xxl	flyest	
festive	freezing	blogger	deadliest	

Figure 5.8. Page 6 of 6.

## *Chapter 6*

---

# International Events from the News

---

(This chapter was originally published as O'Connor et al. (2013).)

## 6.1 Synopsis

The previous chapters described a text analysis tool, as well as opinion and sociolinguistic analyses of social media, that operate at the level of word frequencies, which are a fundamental unit of linguistic representation. However, for analyzing many social questions, the objects of interest may involve more complex objects such as relationships among entities. This chapter specifically looks at textual descriptions of events, which typically engage more complex linguistic structures indicating the semantic relationships between words in a sentence. While more extensive natural language processing techniques are used, similar analysis techniques can be applied—latent variable models, Bayesian inference, and comparisons to previously established social science measurements and data—can be applied.

In this chapter we describe a new probabilistic model for extracting events between major political actors from news corpora. Our unsupervised model brings together familiar components in natural language processing (like parsers and topic models) with contextual political information—temporal and dyad dependence—to infer latent event classes. We quantitatively evaluate the model’s performance on political science benchmarks: recovering expert-assigned event class valences, and detecting real-world conflict. We also conduct a small case study based on our model’s inferences.

## 6.2 Introduction

The digitization of large news corpora has provided an unparalleled opportunity for the systematic study of international relations. Since the mid-1960s political scientists have used political *events data*, records of public micro-level interactions between major political actors of the form “someone does something to someone else” as reported in the open press (Schrodt, 2012), to study the patterns of interactions between political actors and how they evolve over time. Scaling this data effort to modern corpora presents an information extraction challenge: can a structured collection of accurate, politically relevant events between major political actors be extracted automatically and efficiently? And can they be grouped into meaningful event types with a low-dimensional structure useful for further analysis?

We present an unsupervised approach to event extraction, in which political structure and linguistic evidence are combined. A political context model of the relationship between a pair of political actors imposes a prior distribution over types of linguistic events. Our probabilistic model infers latent frames, each a distribution over textual expressions of a kind of event, as well as a representation of the relationship between each political actor pair at each point in time. We use syntactic preprocessing and a logistic normal topic model, including latent temporal smoothing on the political context prior.

We apply the model in a series of comparisons to benchmark datasets in political science. First, we compare the automatically learned verb classes to a pre-existing ontology and hand-crafted verb patterns from TABARI,<sup>1</sup> an open-source and widely used rule-based event extraction system for this domain. Second, we demonstrate correlation to a database of real-world international conflict events, the Militarized Interstate Dispute (MID) dataset (Jones et al., 1996). Third, we qualitatively examine a prominent case not included in the MID dataset, Israeli-Palestinian relations, and compare the recovered trends to the historical record.

We outline the data used for event discovery (§6.3), describe our model (§6.4), inference (§6.5), evaluation (§6.6), and comment on related work (§6.7).

### 6.3 Data

The model we describe in §6.4 is learned from a corpus of 6.5 million newswire articles from the English Gigaword 4th edition (1994–2008, Parker et al., 2009). We also supplement it with a sample of data from the *New York Times* Annotated Corpus (1987–2007, Sandhaus, 2008).<sup>2</sup> The Stanford CoreNLP system,<sup>3</sup> under default settings, was used to POS-tag and parse the articles, to eventually produce event tuples of the form

$$\langle s, r, t, w_{\text{predpath}} \rangle$$

where  $s$  and  $r$  denote “source” and “receiver” arguments, which are political actor entities in a predefined set  $\mathcal{E}$ ,  $t$  is a timestep (i.e., a 7-day period) derived from the article’s published date, and  $w_{\text{predpath}}$  is a textual predicate expressed as a dependency path that typically includes a verb (we use the terms “predicate-path” and “verb-path” interchangeably). For example, on January 1, 2000, the AP reported “Pakistan promptly accused India,” from which our preprocessing extracts the tuple  $\langle \text{PAK}, \text{IND}, 678, \text{accuse} \xleftarrow{\text{dobj}} \rangle$ . (The path excludes the first source-side arc.) Entities and verb paths are identified through the following sets of rules.

Named entity recognition and resolution is done deterministically by finding instances of country names from the *CountryInfo.txt* dictionary from TABARI,<sup>4</sup> which contains proper noun and adjectival forms for countries and administrative units. We supplement these with a few entries for international organizations from another dictionary provided by the same project, and clean up a few ambiguous names, resulting in a final actor dictionary of 235 entities and 2,500 names.

Whenever a name is found, we identify its entity’s mention as the minimal noun phrase that contains it; if the name is an adjectival or noun-noun compound modifier, we traverse any such

---

<sup>1</sup>Available from the Penn State Event Data Project: <http://eventdata.psu.edu/>

<sup>2</sup>For arbitrary reasons this portion of the data is much smaller (we only parse the first five sentences of each article, while Gigaword has all sentences parsed), resulting in less than 2% as many tuples as from the Gigaword data.

<sup>3</sup><http://nlp.stanford.edu/software/corenlp.shtml>

<sup>4</sup><http://eventdata.psu.edu/software.dir/dictionaries.html>.

*amod* and *nn* dependencies to the noun phrase head. Thus *NATO bombing*, *British view*, and *Palestinian militant* resolve to the entity codes IGONAT, GBR, and PSE respectively.

We are interested in identifying actions initiated by agents of one country targeted towards another, and hence concentrate on verbs, analyzing the “CCprocessed” version of the Stanford Dependencies (de Marneffe and Manning, 2008). Verb paths are identified by looking at the shortest dependency path between two mentions in a sentence. If one of the mentions is immediately dominated by a *nsubj* or *agent* relation, we consider that the Source actor, and the other mention is the Receiver. The most common cases are simple direct objects and prepositional arguments like *talk*  $\xleftarrow{\text{prep\_with}}$  and *fight*  $\xleftarrow{\text{prep\_alongside}}$  (“S talk with R,” “S fight alongside R”) but many interesting multiword constructions also result, such as *reject*  $\xleftarrow{\text{dobj}}$  *allegation*  $\xleftarrow{\text{poss}}$  (“S rejected R’s allegation”) or verb chains as in *offer*  $\xleftarrow{\text{xcomp}}$  *help*  $\xleftarrow{\text{dobj}}$  (“S offered to help R”).

We wish to focus on instances of directly reported events, so attempt to remove factively complicated cases such as indirect reporting and hypotheticals by discarding all predicate paths for which any verb on the path has an off-path governing verb with a non-*conj* relation. (For example, the verb at the root of a sentence always survives this filter.) Without this filter, the  $\langle s, r, w \rangle$  tuple  $\langle \text{USA, CUB, want} \xleftarrow{\text{xcomp}} \text{seize} \xleftarrow{\text{dobj}} \rangle$  is extracted from the sentence “Parliament Speaker Ricardo Alarcon said the United States wants to seize Cuba and take over its lands”; the filter removes it since *wants* is dominated by an off-path verb through *say*  $\xleftarrow{\text{ccomp}}$  *wants*. The filter was iteratively developed by inspecting dozens of output examples and their labelings under successive changes to the rules.

Finally, only paths length 4 or less are allowed, the final dependency relation for the receiver may not be *nsubj* or *agent*, and the path may not contain any of the dependency relations *conj*, *parataxis*, *det*, or *dep*. We use lemmatized word forms in defining the paths.

Several document filters are applied before tuple extraction. Deduplication removes 8.5% of articles.<sup>5</sup> For topic filtering, we apply a series of keyword filters to remove sports and finance news, and also apply a text classifier for diplomatic and military news, trained on several hundred manually labeled news articles (using  $\ell_1$ -regularized logistic regression with unigram and bigram features). Other filters remove non-textual junk and non-standard punctuation likely to cause parse errors.

For experiments we remove tuples where the source and receiver entities are the same, and restrict to tuples with dyads that occur at least 500 times, and predicate paths that occur at least 10 times. This yields 365,623 event tuples from 235,830 documents, for 421 dyads and 10,457 unique predicate paths. We define timesteps to be 7-day periods, resulting in 1,149 discrete timesteps (1987 through 2008, though the vast majority of data starts in 1994).

## 6.4 Model

We design two models to learn linguistic event classes over predicate paths by conditioning on real-world contextual information about international politics,  $p(w_{\text{predpath}} | s, r, t)$ , leveraging the fact there tends to be dyadic and temporal coherence in international relations: the types of actions that are likely to occur between nations tend to be similar within the same dyad, and usually their distribution changes smoothly over time.

---

<sup>5</sup>We use a simple form of shingling (ch. 3, Rajaraman and Ullman, 2011): represent a document signature as its  $J = 5$  lowercased bigrams with the lowest hash values, and reject a document with a signature that has been seen before within the same month.  $J$  was manually tuned, as it affects the precision/recall tradeoff.

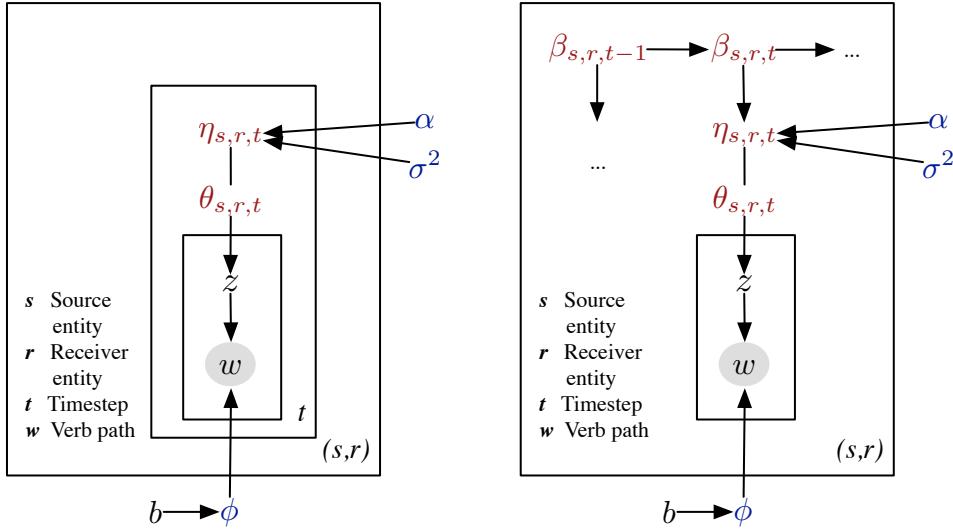


Figure 6.1: Directed probabilistic diagram of the independent timestep model (left) and the smoothed model (right).

Our model decomposes into two submodels: a Context submodel, which encodes how political context affects the probability distribution over event types, and a Language submodel, for how those events are manifested as textual predicate paths (Figure 6.1). The overall generative process is as follows. We color global parameters for a frame **blue**, and local context parameters **red**, and use the term “frame” as a synonym for “event type.” The fixed hyperparameter  $K$  denotes the number of frames.

- The **context model** generates a frame prior  $\theta_{s,r,t}$  for every context  $(s, r, t)$ .
- **Language model:**
  - Draw lexical sparsity parameter  $b$  from a diffuse prior (see §6.10.4).
  - For each frame  $k$ , draw a multinomial distribution of dependency paths,  $\phi_k \sim \text{Dir}(b/V)$  (where  $V$  is the number of dependency path types).
  - For each  $(s, r, t)$ , for every event tuple  $i$  in that context,
    - Sample its frame  $z^{(i)} \sim \text{Mult}(\theta_{s,r,t})$ .
    - Sample its predicate realization  $w_{\text{predpath}}^{(i)} \sim \text{Mult}(\phi_{z^{(i)}})$ .

Thus the language model is very similar to a topic model’s generation of token topics and word-types.

We use structured logistic normal distributions to represent contextual effects. The simplest is the **vanilla** (v) context model,

- For each frame  $k$ , draw global parameters from diffuse priors: prevalence  $\alpha_k$  and variability  $\sigma_k^2$ .
- For each  $(s, r, t)$ ,

- Draw  $\eta_{k,s,r,t} \sim N(\alpha_k, \sigma_k^2)$  for each frame  $k$ .
- Apply a softmax transform,

$$\theta_{k,s,r,t} = \frac{\exp \eta_{k,s,r,t}}{\sum_{k'=1}^K \exp \eta_{k',s,r,t}}$$

Thus the vector  $\eta_{*,s,r,t}$  encodes the relative log-odds of the different frames for events appearing in the context  $(s, r, t)$ . This simple logistic normal prior is, in terms of topic models, analogous to the asymmetric Dirichlet prior version of LDA in [Wallach et al. \(2009\)](#), since the  $\alpha_k$  parameter can learn that some frames tend to be more likely than others. The variance parameters  $\sigma_k^2$  control admixture sparsity, and are analogous to a Dirichlet's concentration parameter.

### 6.4.1 Smoothing Frames Across Time

The vanilla model is capable of inducing frames through dependency path co-occurrences, when multiple events occur in a given context. However, many dyad-time slices are very sparse; for example, most dyads (all but 18) have events in fewer than half the time slices in the dataset. One solution is to increase the bucket size (e.g., to months); however, previous work in political science has demonstrated that answering questions of interest about reciprocity dynamics requires recovering the events at weekly or even daily granularity ([Shellman, 2004](#)), and in any case wide buckets help only so much for dyads with fewer events or less media attention. Therefore we propose a **smoothed frames (SF)** model, in which the frame distribution for a given dyad comes from a latent parameter  $\beta_{*,s,r,t}$  that smoothly varies over time. For each  $(s, r)$ , draw the first timestep's values as  $\beta_{k,s,r,1} \sim N(0, 100)$ , and for each context  $(s, r, t > 1)$ ,

- Draw  $\beta_{k,s,r,t} \sim N(\beta_{k,s,r,t-1}, \tau^2)$
- Draw  $\eta_{k,s,r,t} \sim N(\alpha_k + \beta_{k,s,r,t}, \sigma_k^2)$

Other parameters  $(\alpha_k, \sigma_k^2)$  are same as the vanilla model. This model assumes a random walk process on  $\beta$ , a variable which exists even for contexts that contain no events. Thus inferences about  $\eta$  will be smoothed according to event data at nearby timesteps. This is an instance of a linear Gaussian state-space model (also known as a linear dynamical system or dynamic linear model), and is a convenient formulation because it has well-known exact inference algorithms. Dynamic linear models have been used elsewhere in machine learning and political science to allow latent topic frequencies ([Blei and Lafferty, 2006b](#); [Quinn et al., 2010](#)) and ideological positions ([Martin and Quinn, 2002](#)) to smoothly change over time, and thus share statistical strength between timesteps.

## 6.5 Inference

After randomly initializing all  $\eta_{k,s,r,t}$ , inference is performed by a blocked Gibbs sampler, alternating resamplings for three major groups of variables: the language model  $(z, \phi)$ , context model  $(\alpha, \gamma, \beta, p)$ , and the  $\eta, \theta$  variables, which bottleneck between the submodels.

The **language model** sampler sequentially updates every  $z^{(i)}$  (and implicitly  $\phi$  via collapsing) in the manner of [Griffiths and Steyvers \(2004\)](#):  $p(z^{(i)} | \theta, w^{(i)}, b) \propto \theta_{s,r,t,z} (n_{w,z} + b/V) / (n_z + b)$ , where counts  $n$  are for all event tuples besides  $i$ .

For the **context model**,  $\alpha$  is conjugate resampled as a normal mean. The random walk variables  $\beta$  are sampled with the forward-filtering-backward-sampling algorithm (FFBS; [Harrison and](#)

West, 1997; Carter and Kohn, 1994). There is one slight modification of the standard dynamic linear model that the zero-count weeks have no  $\eta$  observation; the Kalman filter implementation is appropriately modified to handle this.

The  $\eta$  update step is challenging since it is a nonconjugate prior to the  $z$  counts. Logistic normal distributions were introduced to text modeling by Blei and Lafferty (2007), who developed a variational approximation; however, we find that experimenting with different models is easier in the Gibbs sampling framework. While Gibbs sampling for logistic normal priors is possible using auxiliary variable methods (Mimno et al., 2008; Holmes and Held, 2006; Polson et al., 2012), it can be slow to converge. We opt for the more computationally efficient approach of Zeger and Karim (1991) and Hoff (2003), using a Laplace approximation to  $p(\eta \mid \bar{\eta}, \Sigma, z)$ , which is a mode-centered Gaussian having inverse covariance equal to the unnormalized log-posterior's negative Hessian (§8.4 in Murphy, 2012). We find the mode with the linear-time Newton algorithm from Eisenstein et al. (2011b), and sample in linear time by only using the Hessian's diagonal as the inverse covariance (i.e., an axis-aligned normal), since a full multivariate normal sample requires a cubic-time-to-compute Cholesky root of the covariance matrix. This  $\eta^*$  sample is a proposal for a Metropolis-within-Gibbs step, which is moved to according to the standard Metropolis-Hastings acceptance rule. Acceptance rates differ by  $K$ , ranging approximately from 30% ( $K = 100$ ) to nearly 100% (small  $K$ ).

Finally, we use diffuse priors on all global parameters, conjugate resampling variances  $\tau^2, \sigma_k$  once per iteration, and slice sampling (Neal, 2003) the Dirichlet concentration  $b$  every 100 iterations. Automatically learning these was extremely convenient for model-fitting; the only hyper-parameter we set manually was  $K$ . It also allowed us to monitor the convergence of dispersion parameters to help debug and assess MCMC mixing. For other modeling and implementation details, see the appendix section (§6.10).

## 6.6 Experiments

We fit the two models on the dataset described in §6.3, varying the number of frames  $K$ , with 8 or more separate runs for each setting. Posteriors are saved and averaged from 11 Gibbs samples (every 100 iterations from 9,000 to 10,000) for analysis.

We present intrinsic (§6.6.1) and extrinsic (§6.6.2) quantitative evaluations, and a qualitative case study (§6.6.4).

### 6.6.1 Lexical Scale Impurity

In the international relations literature, much of the analysis of text-based events data makes use of a unidimensional conflict to cooperation scale. A popular event ontology in this domain, CAMEO, consists of around 300 different event types, each given an expert-assigned scale in the range from  $-10$  to  $+10$  (Gerner et al., 2002), derived from a judgement collection experiment in Goldstein (1992). The TABARI pattern-based event extraction program comes with a list of almost 16,000 manually engineered verb patterns, each assigned to one CAMEO event type.

It is interesting to consider the extent to which our unsupervised model is able to recover the expert-designed ontology. Given that many of the categories are very fine-grained (e.g. “Express intent to de-escalate military engagement”), we elect to measure model quality as *lexical scale purity*: whether all the predicate paths within one automatically learned frame tend to have similar gold-standard scale scores. (This measures cluster cohesiveness against a one-dimensional continuous scale, instead of measuring cluster cohesiveness against a gold-standard clustering as in VI (Meilă, 2007), Rand index (1971), or cluster purity.) To calculate this, we construct a mapping

between our corpus-derived verb path vocabulary and the TABARI verb patterns, many of which contain one to several word stems that are intended to be matched in surface order. Many of our dependency paths, when traversed from the source to receiver direction, also follow surface order, due to English’s SVO word order.<sup>6</sup> Therefore we convert each path to a word sequence and match against the TABARI lexicon—plus a few modifications for differences in infinitives and stemming—and find 528 dependency path matches. We assign each path  $w$  a gold-standard scale  $g(w)$  by resolving through its matching pattern’s CAMEO code.

We formalize *lexical scale impurity* as the average absolute difference of scale values between two predicate paths under the same frame. Specifically, we want a token-level posterior expectation

$$\mathbb{E}(|g(w_i) - g(w_j)| \mid z_i = z_j, w_i \neq w_j) \quad (6.1)$$

which is taken over pairs of path instances  $(i, j)$  where both paths  $w_i, w_j$  are in  $M$ , the set of verb paths that were matched between the lexicons. This can be reformulated at the type level as:<sup>7</sup>

$$\frac{1}{N} \sum_k \sum_{\substack{w, v \in M \\ w \neq v}} n_{w,k} n_{v,k} |g(w) - g(v)| \quad (6.2)$$

where  $n$  refers to the averaged Gibbs samples’ counts of event tuples having frame  $k$  and a particular verb path,<sup>8</sup> and  $N$  is the number of token comparisons (i.e. the same sum, but with a 1 replacing the distance). The worst possible impurity is upper bounded at 20 ( $= \max(g(w)) - \min(g(w))$ ) and the best possible is 0. We also compute a randomized null hypothesis to see how low impurity can be by chance: each of  $\sim 1000$  simulations randomly assigns each path in  $M$  to one of  $K$  frames (all its instances are exclusively assigned to that frame), and computes the impurity. On average the impurity is same at all  $K$ , but variance increases with  $K$  (since small clusters might by chance get a highly similar paths in them), necessitating this null hypothesis analysis. We report the 5th percentile over simulations.

## 6.6.2 Conflict Detection

Political events data has shown considerable promise as a tool for crisis early warning systems (O’Brien, 2010; Brandt et al., 2011). While conflict forecasting is a potential application of our model, we conduct a simpler prediction task to validate whether the model is learning something useful: based on news text, tell whether or not an armed conflict is *currently* happening. For a gold standard, we use the Militarized Interstate Dispute (MID) dataset (Jones et al., 1996; Ghosn et al., 2004), which documents historical international disputes. While not without critics, the MID data is the most prominent dataset in the field of international relations. We use the Dyadic MIDs, each of which ranks hostility levels between pairs of actors on a five point scale over a date interval; we define conflict to be the top two categories “Use of Force” (4) and “War” (5). We convert the data into a variable  $y_{s,r,t}$ , the highest hostility level reached by actor  $s$  directed towards receiver  $r$  in the dispute that overlaps with our 7-day interval  $t$ , and want to predict the binary indicator  $\mathbf{1}\{y_{s,r,t} \geq 4\}$ . For the illustrative examples (USA to Iraq, and the Israel-Palestine example below) we use results from a smaller but more internally comparable dataset consisting of the 2 million Associated Press articles within the Gigaword corpus.

<sup>6</sup>There are exceptions where a Source-to-Receiver path traversal can have a right-to-left move, such as dependency edges for possessives, in e.g.  $S \xrightarrow{\text{nssubj}} \text{partner} \xleftarrow{\text{poss}} R$ , as in “ $S$  is  $R$ ’s partner”. This approach can not match them.

<sup>7</sup>Derivation in §6.9.1.

<sup>8</sup>Results are nearly identical whether we use counts averaged across samples (thus giving posterior marginals), or simply use counts from a single sample (i.e., iteration 10,000).

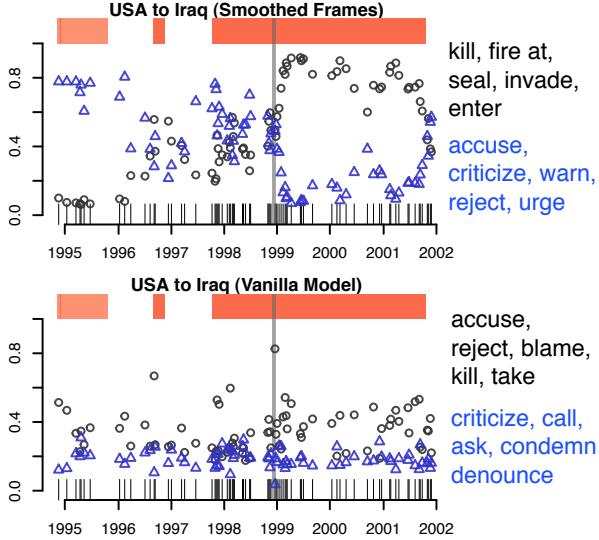


Figure 6.2: The USA→Iraq directed dyad, analyzed by smoothed (above) and vanilla (below) models, showing (1) gold-standard MID values (red intervals along top), (2) weeks with non-zero event counts (vertical lines along x-axis), (3) posterior  $E[\theta_k, \text{USA}, \text{IRQ}, t]$  inferences for two frames chosen from two different  $K = 5$  models, and (4) most common verb paths for each frame (right). Frames corresponding to material and verbal conflict were chosen for display. Vertical line indicates Operation Desert Fox (see §6.6.2).

For an example of the MID data, see Figure 6.2, which depicts three disputes between the US and Iraq in this time period. The MID labels are marked in red.

The first dispute is a “display of force” (level 3), cataloguing the U.S. response to a series of troop movements along the border with Kuwait. The third dispute (10/7/1997 to 10/10/2001) begins with increasing Iraqi violations of the no-fly zone, resulting in U.S. and U.K. retaliation, reaching a high intensity with Operation Desert Fox, a four-day bombing campaign from December 16 to 19, 1998—which is not shown in MID. These cases highlight MID’s limitations—while it is well regarded in the political science literature, its coarse level of aggregation can fail to capture variation in conflict intensity.

Figure 6.2 also shows model inferences. Our smoothed model captures some of these phenomena here, showing clear trends for two relevant frames, including a dramatic change in December 1998. The vanilla model has a harder time, since it cannot combine evidence between different timesteps.

The MID dataset overlaps with our data for 470 weeks, from 1993 through 2001. After excluding dyads with actors that the MID data does not intend to include—Kosovo, Tibet, Palestine, and international organizations—we have 267 directed dyads for evaluation, 117 of which have at least one dispute in the MID data. (Dyads with no dispute in the MID data, such as Germany-France, are assumed to have  $y = 0$  throughout the time period.) About 7% of the dyad-time contexts have a dispute under these definitions.

We split the dataset by time, training on the first half of the data and testing on the second half, and measure area under the receiver operating characteristic curve (AUC).<sup>9</sup> For each model, we

<sup>9</sup>AUC can be interpreted as follows: given a positive and negative example, what is the probability that the classifier’s confidences order them correctly? Random noise or predicting all the same class both give AUC 0.5.

train an  $\ell_1$ -regularized logistic regression<sup>10</sup> with the  $K$  elements of  $\theta_{*,s,r,t}$  as input features, tuning the regularization parameter within the training set (by splitting it in half again) to optimize held-out likelihood. We weight instances to balance positive and negative examples. Training is on all individual  $\theta$  samples at once (thus accounting for posterior uncertainty in learning), and final predicted probabilities are averaged from individual probabilities from each  $\theta$  test set sample, thus propagating posterior uncertainty into the predictions. We also create a baseline  $\ell_1$ -regularized logistic regression that uses normalized dependency path counts as the features (10,457 features). For both the baseline and vanilla model, contexts with no events are given a feature vector of all zeros.<sup>11</sup> (We also explored an alternative evaluation setup, to hold out by dyad; however, the performance variance is quite high between different random dyad splits.)

### 6.6.3 Results

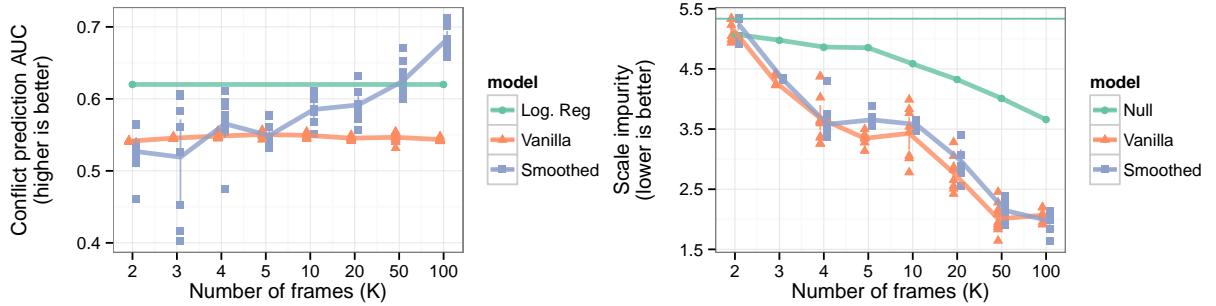


Figure 6.3: Evaluation results. Each point indicates one model run. Lines show the average per  $K$ , with vertical lines indicating the 95% bootstrapped interval. **Left:** Conflict detection AUC for different models (§6.6.2). Green line is the verb-path logistic regression baseline. **Right:** Lexical scale impurity (§6.6.1). Top green line indicates the simple random baseline  $E(|g(w_i) - g(w_j)|) = 5.33$ ; the second green line is from the random assignment baseline.

Results are shown in Figure 6.3.<sup>12</sup>

The verb-path logistic regression performs strongly at AUC 0.62; it outperforms all of the vanilla frame models. This is an example of individual lexical features outperforming a topic model for predictive task, because the topic model’s dimension reduction obscures important indicators from individual words. Similarly, [Gerrish and Blei \(2011\)](#) found that word-based regression outperformed a customized topic model when predicting Congressional bill passage, and Chapter 4 found word-based regression outperformed Supervised LDA for geolocation, and we have noticed this phenomenon for other text-based prediction problems.

However, adding smoothing to the model substantially increases performance, and in fact outperforms the verb-path regression at  $K = 100$ . It is unclear why the vanilla model fails to increase performance in  $K$ . Note also, the vanilla model exhibits very little variability in prediction performance between model runs, in comparison to the smoothed model which is much more variable (presumably due to the higher number of parameters in the model); at small values of  $K$ , the smoothed model can perform poorly. It would also be interesting to analyze the smoothed model with higher values of  $K$  and find where it peaks.

<sup>10</sup>Using the R `glmnet` package ([Friedman et al., 2010](#)).

<sup>11</sup>For the vanilla model, this performed better than linear interpolation (about 0.03 AUC), and with less variance between runs.

<sup>12</sup>Due to an implementation bug, the model put the vast majority of the probability mass only on  $K - 1$  frames, so these settings might be better thought of as  $K = 1, 2, 3, 4, 9, \dots$ ; see the appendix for details.

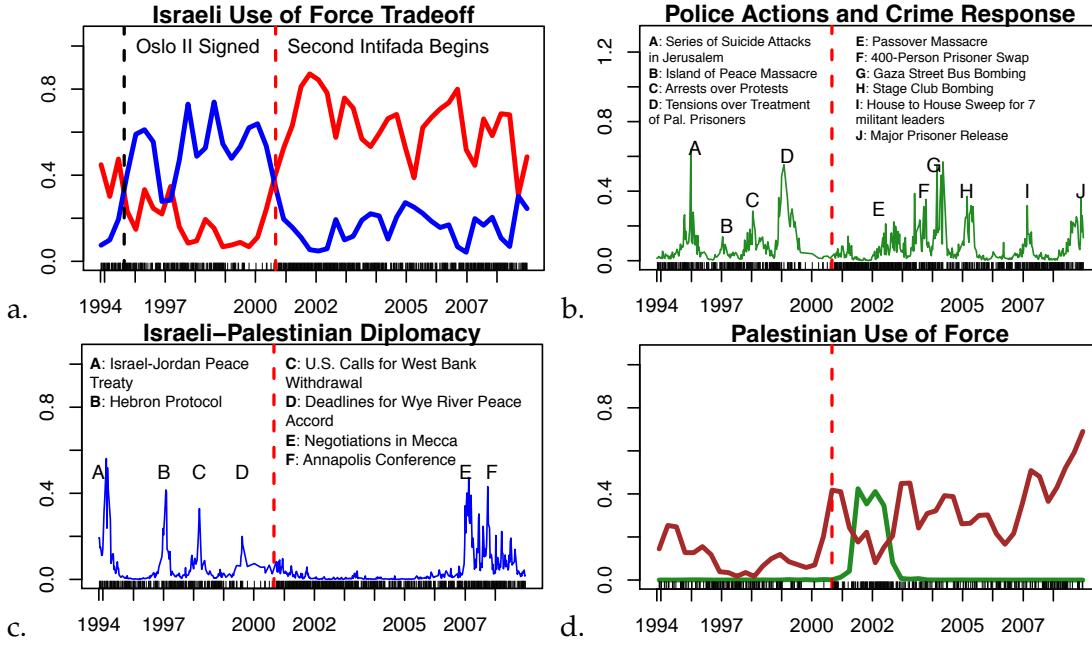


Figure 6.4: For Israel-Palestinian directed dyads, plots of  $E[\theta]$  (proportion of weekly events in a frame) over time, annotated with historical events. (a): Words are ‘kill, fire at, enter, kill, attack, raid, strike, move, pound, bomb’ and ‘impose, seal, capture, seize, arrest, ease, close, deport, close, release’ (b): ‘accuse, criticize, reject, tell, hand to, warn, ask, detain, release, order’ (c): ‘meet with, sign with, praise, say with, arrive in, host, tell, welcome, join, thank’ (d): again the same ‘kill, fire at’ frame in (a), plus the erroneous frame (see text) ‘include, join, fly to, have relation with, protest to, call, include bomber  $\xleftarrow{\text{appos}}$  informer for’. Figures (b) and (c) use linear interpolation for zero-count weeks (thus relying exclusively on the model for smoothing); (a) and (d) apply a lowess smoother. (a-c) are for the ISR→PSE direction; (d) is PSE→ISR.

We view the conflict detection task only as one of several validations, and thus turn to lexical evaluation of the induced frames. For lexical scale purity (right side of Figure 6.3), the models perform about the same, with the smoothed model a little bit worse at some values of  $K$  (though sometimes with better stability of the fits—opposite of the conflict detection task). This suggests that semantic coherence does not benefit from the longer-range temporal dependencies.

In general, performance improves with higher  $K$ , but not beyond  $K = 50$ . This suggests the model reaches a limit for how fine-grained of semantics it can learn.

#### 6.6.4 Case study

Here we qualitatively examine the narrative story between the dyad with the highest frequency of events in our dataset, the Israeli-Palestinian relationship, finding qualitative agreement with other case studies of this conflict (Brandt et al., 2012; Goldstein et al., 2001; Schrot and Gerner, 2004). (The MID dataset does not include this conflict because the Palestinians are not considered a state actor.) Using the Associated Press subset, we plot the highest incidence frames from one run of the  $K = 20$  smoothed frame models, for the two directed dyads, and highlight some of the interesting relationships.

Figure 6.4(a) shows that tradeoffs in the use of military vs. police action by Israel towards the

Palestinians tracks with major historical events. The first period in the data where police actions ('impose, seal, capture, seize, arrest') exceed military actions ('kill, fire, enter, attack, raid') is with the signing of the "Interim Agreement on the West Bank and the Gaza Strip," also known as the Oslo II agreement. This balance persists until the abrupt breakdown in relations that followed the unsuccessful Camp David Summit in July of 2000, which generally marks the starting point of the wave of violence known as the Second Intifada.

In Figure 6.4(b) we show that our model produces a frame which captures the legal aftermath of particular events ('accuse, criticize,' but also 'detain, release, extradite, charge'). Each of the major spikes in the data coincides with a particular event which either involves the investigation of a particular attack or series of attacks (as in A,B,E) or a discussion about prisoner swaps or mass arrests (as in events D, F, J).

Our model also picks up positive diplomatic events, as seen in Figure 6.4(c), a frame describing Israeli diplomatic actions towards Palestine ('meet with, sign with, praise, say with, arrive in'). Not only do the spikes coincide with major peace treaties and negotiations, but the model correctly characterizes the relative lack of positively valenced action from the beginning of the Second Intifada until its end around 2005–2006.

In Figure 6.4(d) we show the relevant frames depicting use of force from the Palestinians towards the Israelis (brown trend line). At first, the drop in the use of force frame immediately following the start of the Second Intifada seems inconsistent with the historical record. However, there is a concurrent rise in a different frame driven by the word 'include', which actually appears here due to an NLP error compounded with an artifact of the data source. A casualties report article, containing variants of the text "The Palestinian figure includes... 13 Israeli Arabs...", is repeated 27 times over two years. "Palestinian figure" is erroneously identified as the PSE entity, and several noun phrases in a list are identified as separate receivers. This issue causes 39 of all 86 PSE→ISR events during this period to use the word 'include', accounting for the rise in that frame. (This highlights how better natural language processing could help the model, and the dangers of false positives for this type of data analysis, especially in small-sample drilldowns.) Discounting this erroneous inference, the results are consistent with heightened violence during this period.

We conclude that the frame extractions for the Israeli-Palestinian case are consistent with the historical record over the period of study.

## 6.7 Related Work

### 6.7.1 Events Data in Political Science

Projects using hand-collected events data represent some of the earliest efforts in the statistical study of international relations, dating back to the 1960s (Rummel, 1968; Azar and Sloan, 1975; McClelland, 1970). Beginning in the mid-1980s, political scientists began experimenting with automated rule-based extraction systems (Schrodt and Gerner, 1994). These efforts culminated in the open-source program, TABARI, which uses pattern matching from extensive hand-developed phrase dictionaries, combined with basic part of speech tagging (Schrodt, 2001); a rough analogue in the information extraction literature might be the rule-based, finite-state FASTUS system for MUC IE (Hobbs et al., 1997), though TABARI is restricted to single sentence analysis. Later proprietary work has apparently incorporated more extensive NLP (e.g., sentence parsing) though few details are available (King and Lowe, 2003). The most recent published work we know of, by Boschee et al. (2013), uses a proprietary parsing and coreference system (BBN SERIF, Ramshaw et al., 2011), and directly compares to TABARI, finding significantly higher accuracy. The original TABARI system is still actively being developed, including just-released work on a new 200

million event dataset, GDELT (Schrodt and Leetaru, 2013).<sup>13</sup> All these systems crucially rely on hand-built pattern dictionaries.

It is extremely labor intensive to develop these dictionaries. Schrodt (2006) estimates 4,000 trained person-hours were required to create dictionaries of political actors in the Middle East, and the phrase dictionary may have taken even longer; the comments in TABARI’s phrase dictionary indicate some of its 15,789 entries were created as early as 1991. Ideally, any new events data solution would incorporate the extensive work already completed by political scientists in this area while minimizing the need for further dictionary development. In this work we use the actor dictionaries, and hope to incorporate the verb patterns in future work.

### 6.7.2 Events in Natural Language Processing

Political event extraction from news has also received considerable attention within natural language processing in part due to research programs such as MUC-3 and MUC-4 (Lehnert, 1994), which focused on the extraction of terrorist events, as well as the more recent ACE program. The work in this paper is inspired by unsupervised approaches that seek to discover types of relations and events, instead of assuming them to be pre-specified; this includes research under various headings such as template/frame/event learning (Cheung et al., 2013; Modi et al., 2012; Chambers and Jurafsky, 2011; Li et al., 2010; Bejan, 2008), script learning (Regneri et al., 2010; Chambers and Jurafsky, 2009), relation learning (Yao et al., 2011), open information extraction (Banko et al., 2007; Carlson et al., 2010), verb caseframe learning (Rooth et al., 1999; Gildea, 2002; Grenager and Manning, 2006; Lang and Lapata, 2010; Ó Séaghdha, 2010; Titov and Klementiev, 2012), and a version of frame learning called “unsupervised semantic parsing” (Titov and Klementiev, 2011; Poon and Domingos, 2009). Unlike much of the previous literature, we do not learn latent roles/slots. Event extraction is also a large literature, including supervised systems targeting problems similar to MUC and political events (Piskorski and Atkinson, 2011; Piskorski et al., 2011; Sanfilippo et al., 2008).

One can also see this work as a relational extension of co-occurrence-based methods such as Gerrish (2013; ch. 4), Diesner and Carley (2005), Chang et al. (2009), or Newman et al. (2006), which perform bag-of-words-style analysis of text fragments containing co-occurring entities. (Gerrish also analyzed the international relations domain, using supervised bag-of-words regression to assess the expressed valence between a pair of actors in a news paragraph, using the predictions as observations in a latent temporal model, and compared to MID.) We instead use parsing to get a much more focused and interpretable representation of the relationship between textually co-occurring entities; namely, that they are the source and target of an action event. This is more in line with work in relation extraction on biomedical scientific articles (Friedman et al., 2001; Rzhetsky et al., 2004) which uses parsing to extracting a network of how different entities, like drugs or proteins, interact.

## 6.8 Conclusion

Large-scale information extraction can dramatically enhance the study of political behavior. Here we present a novel unsupervised approach to an important data collection effort in the social sciences. We see international relations as a rich and practically useful domain for the development of text analysis methods that jointly infer events, relations, and sociopolitical context. There are numerous areas for future work, such as: using verb dictionaries as semi-supervised seeds or priors;

---

<sup>13</sup><http://eventdata.psu.edu/data.dir/GDELT.html>

interactive learning between political science researchers and unsupervised algorithms; building low-dimensional scaling, or hierarchical structure, into the model; and learning the actor lists to handle changing real-world situations and new domains. In particular, adding more supervision to the model will be crucial to improve semantic quality and make it useful for researchers.

## 6.9 Appendix: Evaluation

All references to software and data files are for the materials available at <http://brenocon.com/irevents/>. File names below are linked to URLs, but should be usable with any downloaded version of the software and data.

### 6.9.1 Type-level calculation of lexical scale impurity

As noted in §6.6.1, the measure we want is a posterior expectation defined for instance pairs, which we can reformulate at the type level as follows. Let  $i$  and  $j$  index over instances, and  $w$  and  $v$  index over types. Consider an expectation using a single sample to represent the posterior,

$$E [ |g(w_i) - g(w_j)| \mid z_i = z_j \& w_i \neq w_j \& w_i, w_j \in M] = \frac{Q}{N} \quad (6.3)$$

where  $N$  is the number of instance pair comparisons satisfying the conditional, and  $Q$  is,

$$Q = \sum_{ij} 1\{z_i = z_j\} 1\{w_i \in M\} 1\{w_j \in M\} 1\{w_i \neq w_j\} d_{ij} \quad (6.4)$$

$$= \sum_k \sum_{ij} 1\{z_i = k\} 1\{z_j = k\} 1\{w_i \in M\} 1\{w_j \in M\} 1\{w_i \neq w_j\} d_{ij} \quad (6.5)$$

$$= \sum_k \sum_i 1\{z_i = k\} 1\{w_i \in M\} \sum_j 1\{z_j = k\} 1\{w_j \in M\} 1\{w_i \neq w_j\} d_{ij} \quad (6.6)$$

$$= \sum_k \sum_{w \in M, v \in M, w \neq v} n_{wk} n_{vk} d_{wv} \quad (6.7)$$

where  $d_{ij} = |g(w_i) - g(w_j)|$ ,  $d_{wv} = |g(w) - g(v)|$ , and  $n_{wk}$  and  $n_{vk}$  are from the collapsed Gibbs sampling count tables, i.e.  $n_{wk} = \sum_i 1\{w_i = w\} 1\{z_i = k\}$ .

The denominator is

$$N = \sum_k \sum_{w \in M, v \in M, w \neq v} n_{w,k} n_{v,k}$$

To properly compute a posterior expectation using multiple samples,  $Q/N$  should be re-evaluated on several complete samples and then averaged. However, we found little variation between samples so used only one. We also tried evaluating a single  $Q/N$  where  $n_{wk}$  and  $n_{vk}$  are *averaged* counts from multiple samples—using this corresponds to a factored, mean-field-like approximation to the posterior—but it also was very similar to using a single sample.

The implementation is in [verbdict/score.py](#).

### 6.9.2 TABARI lexicon matching

Two additional notes:

1. There were a number of patterns in the TABARI lexicon that had multiple conflicting codes. See [verbdict/contradictory\\_codes.txt](#).
2. As described in the paper, the dependency paths are traversed from source to receiver, creating the corresponding word sequence. Prepositions are un-collapsed and put into the sequence. There is special handling of *xcomp*'s, which sometimes represent an infinitival 'to' and sometimes do not; we generate two versions, with and without 'to'; if either one matches to a TABARI pattern then that counts as a match.

The implementation is in [verbdict/match.py](#).

## 6.10 Appendix: Details on Inference

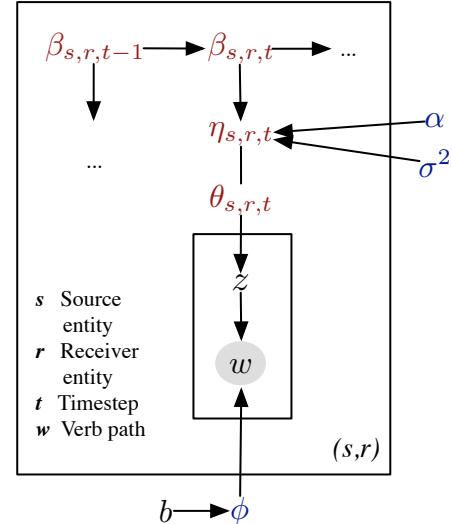
The full smoothed model is:

Context model (smoothed frames):

$$\begin{aligned}\tau^2 &\sim \text{InvGamma} \\ \sigma_k^2 &\sim \text{InvGamma} \\ \alpha_k &\sim \text{Normal} \\ \beta_{s,r,1,k} &\sim N(0, 100) \\ \beta_{s,r,t>1,k} &\sim N(\beta_{k,s,r,t-1}, \tau^2) \\ \eta_{s,r,t,k} &\sim N(\alpha_k + \beta_{k,s,r,t}, \sigma_k^2) \\ \theta_{s,r,t,*} &= \text{Softmax}(\eta_{s,r,t,*})\end{aligned}$$

Language model:

$$\begin{aligned}b &\sim \text{ImproperUniform} \\ \phi_k &\sim \text{Dir}(b/V) \\ z &\sim \theta_{s,r,t} \\ w &\sim \phi_z\end{aligned}$$



The blocked Gibbs sampler proceeds on the following groups of variables. These conditionals also implicitly condition on  $w, s, r, t$ .

- Context (Politics) submodel
  - $[\alpha | \eta, \beta, \sigma^2]$ : Exact
  - $[\beta | \eta, \alpha, \sigma^2]$ : Exact, FFBS algorithm
- Context/Language bridge
  - $[\eta | \beta, \alpha, z]$ : Laplace approximation Metropolis-within-Gibbs step
- Language submodel
  - $[z | \eta]$ : Exact, collapsing out  $\phi$

- Dispersions (variances and concentrations)

- $[\tau^2 | \beta], [\sigma^2 | \eta, \alpha], [b | z, w]$

The key step is sampling instantiations of  $\eta$ , which is the bottleneck between the politics and language models; given that, inference proceeds on either side of the model via well-known conjugate posterior resampling updates, each described as follows.

### 6.10.1 Language Submodel $[z | \eta]$

This is the most straightforward step in light of previous work in Bayesian language modeling. Dirichlet-Multinomial conjugacy allows Gibbs sampling to proceed on individual  $z$ 's for individual tuples, collapsing out  $\phi$  (as in Griffiths and Steyvers (2004), though unlike that work we condition on  $\theta$ ):

$$p(z_i = k | s, r, t, w, z_{-i}, \theta, b) \propto \theta_{s,r,t} \frac{\#\{z = k, w\} + b/V}{\#\{z = k\} + b} \quad (6.8)$$

where the counts are taken from the current  $z$  setting in all corpus tuples, except tuple  $i$ .  $b$  is the Dirichlet concentration parameter, and  $V$  is the number of verb-path types.

### 6.10.2 Context Submodel $[\alpha, \beta | \eta]$

The  $\alpha$  update is just a conjugate normal sample; see any standard Bayesian reference, e.g. §4.4.2.1 of Murphy (2012), or Gelman et al. (2003). Let the all-but- $\alpha$  residual be  $r_{s,r,t,k} = \eta_{s,r,t,k} - \beta_{s,r,t,k}$ , so  $r \sim N(\alpha, \sigma_k^2)$ . With prior  $p(\alpha) \sim N(0, 100)$ , then

$$p(\alpha_k | \eta, \beta, \sigma_k^2) = N\left(\frac{n/\sigma_k^2}{n/\sigma_k^2 + 1/100} \bar{r}_k, [1/100 + n/\sigma_k^2]^{-1}\right)$$

where  $\bar{r}_k$  is the current residual empirical mean:  $\bar{r}_k = \sum_{s,r,t} (\eta_{s,r,t,k} - \beta_{s,r,t,k})$ , and  $n$  is the number of  $\eta$  emissions for this frame (i.e. the number contexts).  $\eta$  only exists for contexts with at least one event tuple (otherwise it is vacuous variable), the sums over  $(s, r, t)$  are only over those contexts. Still,  $n$  is very large (hundreds of thousands) so the posterior is very peaked; updating  $\alpha$  is basically the same as an maximum likelihood estimate and the prior is irrelevant.

The  $\beta$  update is dynamic linear model inference. Because of the emissions' diagonal covariance  $diag(\sigma_1^2 \dots \sigma_K^2)$ , it decomposes into conditional independence for each frame's time series for each dyad. A single joint sample of one of these time series,

$$(\hat{\beta}_{s,r,1,k} \dots \hat{\beta}_{s,r,T,k}) \sim p(\beta_{s,r,1,k} \dots \beta_{s,r,T,k} | \alpha, \eta, \sigma_k^2, \tau^2)$$

can be drawn exactly with dynamic programming, via the forward filter, backward sampling algorithm (FFBS; Harrison and West, 1997; Carter and Kohn, 1994). We leave out  $\alpha, \sigma_k^2, \tau^2$  in the following equations for clarity. Here, FFBS proceeds in two steps: (1) run a Kalman filter, successively computing each  $p(\beta_t | \eta_1 \dots \eta_t)$  (each of which is normal), and (2) run a sampling variant of the RTS smoother, to sample successively each  $\hat{\beta}_t \sim p(\beta_t | \hat{\beta}_{t+1}, \eta_1 \dots \eta_t)$  (each of which is also normal). The final sequence of sampled  $\hat{\beta}_t$  values is a sample from the joint sequence posterior, since  $p(\beta_1 \dots \beta_T | \eta_{1:T}) = p(\beta_T | \eta_{1:T}) p(\beta_{T-1} | \beta_T, \eta_{1:T-1}) \dots p(\beta_1 | \beta_2, \eta_1)$ .

We use  $\mu$  and  $\Sigma$  to denote posterior beliefs about  $\beta$ . Let  $N(\mu_{t|t-1}, \Sigma_{t|t-1})$  denote  $p(\beta_t | \eta_1 \dots \eta_{t-1})$ , and  $N(\mu_t, \Sigma_t)$  denote  $p(\beta_t | \eta_1 \dots \eta_t)$  (where  $\Sigma$  is just a scalar variance). The full Kalman filter is defined for much more general Gaussian state-space models: (Murphy, 2012 §18.3 notation)

$$\begin{aligned} z_t &= Az_{t-1} + Bu_t + N(0, Q) \\ y_t &= Cz_t + Dut + N(0, R) \end{aligned}$$

Which for us is just

$$\begin{aligned} \beta_t &= \beta_{t-1} + N(0, \tau^2) \\ \eta_t &= \beta_t + \alpha + N(0, \sigma^2) \end{aligned}$$

The algorithm is<sup>14</sup>

- Filter, which takes  $\eta_{1:T}$  as input.

Initialize  $\mu_{1|0} := 0$ ,  $\Sigma_{1|0} := 100$  (and skip the prediction step on the first iteration).

For  $t = 1..T$ ,

- Prediction step (infer  $p(\beta_t | \eta_1 \dots \eta_{t-1})$ ):

$$\begin{aligned} \mu_{t|t-1} &:= \mu_{t-1} \\ \Sigma_{t|t-1} &:= \Sigma_{t-1} + \tau^2 \end{aligned}$$

- Measurement step (infer  $p(\beta_t | \eta_1 \dots \eta_t)$ ):

$$\begin{aligned} r &:= \eta_t - (\mu_{t|t-1} + \alpha) \quad (\text{residual}) \\ K &:= \Sigma_{t|t-1} (\Sigma_{t|t-1} + \sigma^2)^{-1} \quad (\text{Kalman gain}) \\ \mu_t &:= \mu_{t|t-1} + Kr \\ \Sigma_t &:= \Sigma_{t|t-1} (1 - K) \end{aligned}$$

- Backward-sampler, which uses the filtered quantities  $\mu_t, \Sigma_t$  as input.

Initially sample  $\hat{\beta}_T \sim N(\mu_T, \Sigma_T)$ .

For  $t = (T-1)..1$ ,

- Sample  $\hat{\beta}_t \sim N(\mu_t + L(\hat{\beta}_{t+1} - \mu_{t+1|t}), \Sigma_t - L^2 \Sigma_{t+1|t})$   
where  $L = \Sigma_t (\Sigma_{t+1|t})^{-1}$

We have one modification to the standard linear dynamical system model: while a  $\beta$  exists for all timesteps, there are many zero-count contexts without any event tuples. The Kalman filter is modified to skip the measurement step for those timesteps, so simply  $\mu_t := \mu_{t|t-1}$  and  $\Sigma_t := \Sigma_{t|t-1}$ . We do not store  $\eta$  variables at those timesteps, since they are unnecessary for inference; but we do simulate them when creating posterior samples for analysis in the conflict detection task. (The time-series plots of  $E[\theta]$  in section 5 of the paper do not show these samples.)

We use a custom implementation of the filter and sampler that was tested via simulation in two ways: (1) comparing its inferences on simulated data to those from the *dlm* package in R (Petris, 2010), and (2) using the Cook et al. (2006) Bayesian software validation technique of checking the simulation distribution of inferred posterior quantiles of simulated parameters. The latter was useful for testing other samplers as well. In fact, the softmax bug in the logistic normal inference procedure (§6.10.5) resulted from a case where the validation tests were not used.

---

<sup>14</sup>See also [http://www.gatsby.ucl.ac.uk/~turner/Notes/1DKalmanFilter/1d\\_kalman\\_filter.pdf](http://www.gatsby.ucl.ac.uk/~turner/Notes/1DKalmanFilter/1d_kalman_filter.pdf).

### 6.10.3 Logistic Normal $[\eta \mid z, \bar{\eta}]$

Next, we must resample the  $\eta$  variables; for every context, sample from the posterior density

$$p(\eta \mid \bar{\eta}, z) \propto N(\eta \mid \bar{\eta}, \Sigma) \text{ } Mult(z \mid \theta(\eta)) \quad (6.9)$$

where  $\bar{\eta} = \beta + \alpha$  denotes  $\eta$ 's prior mean. This has an unnormalized log posterior density function

$$\ell(\eta) = \sum_k \left( -\frac{1}{2\sigma_k^2} (\eta_k - \bar{\eta}_k)^2 + n_k \log \theta(\eta)_k \right) \quad (6.10)$$

where  $n_k$  is the number of tuples in this context having frame  $k$ , and  $\theta(\eta)$  is the value of  $\theta$  deterministically associated with  $\eta$  via the softmax function.

Unfortunately, unlike the Dirichlet, a logistic normal prior on a multinomial is not conjugate; Equation 6.10 describes the unnormalized density, but there is no closed form for the normalized posterior (and more to the point, no known exact sampling algorithm).

As described in the paper, we use a Laplace approximation proposal—a Gaussian approximation centered at the mode, which can be justified as the second-order approximation to the log-posterior there—taking a proposed sample  $\eta^*$  via the steps

- (1) Solve MAP  $\hat{\eta} = \arg \max_{\eta} \ell(\eta)$
- (2) Sample  $\eta^* \sim N(\hat{\eta}, [H(-\ell(\hat{\eta}))]^{-1})$

where  $H(-\ell(\hat{\eta}))$  denotes Hessian of the negative unnormalized log-posterior at  $\hat{\eta}$ .

Step #1 could be solved in a number of ways. We use a fast linear-time Newton algorithm from Eisenstein et al. (2011b), which was faster than gradient descent methods we tried; we reproduce it below. The Newton step is

$$\eta := \eta - \lambda H^{-1} g$$

where the gradient of  $-\ell$  is

$$g(\eta)_k = n\theta_k - n_k + \frac{1}{\sigma_k^2} (\eta_k - \bar{\eta}_k)$$

and the Hessian has diagonal and off-diagonal elements

$$H_{kk} = n\theta_k(1 - \theta_k) + 1/\sigma_k^2, \quad H_{jk} = -n\theta_j\theta_k$$

where  $n$  is the number of event tuples in the context (i.e. number of individual  $z$ 's). Matrix inversion is in general a cubic time algorithm, but we apply the Sherman-Morrison formula to only have to invert a diagonal matrix. For any invertible square matrix  $A$  and vectors  $u, v$ , the Sherman-Morrison formula gives an alternate expression for  $(A + uv^T)^{-1}$  in terms of  $A^{-1}$ . For a diagonal matrix  $A$  and vectors  $u, v, w$ , we apply the Sherman-Morrison formula and configure the order of operations to avoid creating any non-diagonal matrices:

$$Z = (A + uv^T)^{-1}w \quad (6.11)$$

$$Z = A^{-1}w - [1 + v^T A^{-1}u]^{-1}(A^{-1}u)(v^T A^{-1}w) \quad (6.12)$$

$$Z_j = (A_{jj}^{-1}w_j) - \frac{1}{1 + \sum_k A_{kk}^{-1}v_k u_k} (A_{jj}^{-1}u_j) \sum_k A_{kk}^{-1}v_k w_k \quad (6.13)$$

where the last line shows the resulting vector for one element  $j$ .

The Hessian can be rewritten as a sum of diagonal and rank-1 matrix as  $H = \text{diag}[n\theta_k + 1/\sigma_k^2] - n\theta\theta^T$ , thus the Newton step direction  $H^{-1}g$  can be calculated in linear time by applying Eq. 6.12 with  $A_{kk}^{-1} = (n\theta_k + 1/\sigma_k^2)^{-1}$ ,  $w = g$ ,  $u = \sqrt{n}\theta$ ,  $v = -\sqrt{n}\theta$ .

Eisenstein et al. (2011a) present this technique in the context of a variational inference algorithm, but actually it applies to any MAP logistic normal inference problem under diagonal covariance. We find it usually converges to an  $\hat{\eta}$  estimate in only several iterations (using a line search,<sup>15</sup> first taking a step sized  $\lambda = 1$ , and if it's not an improvement, halving  $\lambda$  until it is.)

Step #2 is to sample from the multivariate normal  $N(\hat{\eta}, H^{-1})$ . The simplest MVN sampling algorithm is to take  $K$  samples from  $N(0, 1)$  and multiply that vector by the Cholesky root of the covariance (and add the mean). But it takes cubic time to compute a Cholesky root (in the general case), which is too expensive for large values of  $K$ . Instead, we only invert the diagonal of the Hessian (linear time), resulting in a diagonal covariance (thus each  $\eta_k^* \sim N(\bar{\eta}_k, 1/H_{kk})$ ); this is only an axis-aligned MVN approximation to the posterior.<sup>16</sup>

So this gives a  $\eta^{\text{new}}$  proposal. It is possible to simply update to it directly; but it is more accurate to use it as a Metropolis-Hastings proposal. Calculate the acceptance probability

$$a = \frac{p(\eta^{\text{new}} | \bar{\eta}, z)}{p(\eta^{\text{old}} | \bar{\eta}, z)} \frac{N(\eta^{\text{old}}, \hat{\eta}, H^{-1})}{N(\eta^{\text{new}}, \hat{\eta}, H^{-1})}$$

and accept the proposal at probability  $\min(a, 1)$ . The ratio of true posterior densities can be calculated with the unnormalized form in Equation 6.10.

See also Wang and Blei (2012) which develops a Laplace approximation for variational inference for several nonconjugate models, including a logistic normal topic model; this approach is also applied in Roberts et al. (2013). The Metropolis-Hastings approach we use here is similar to Hoff (2003).

#### 6.10.4 Learning concentrations and variances

There are several parameters that control the overall variability of the above quantities. The Dirichlet concentration parameter  $b$  controls the similarity between the frames' predicate-path distributions; the autoregressive variance  $\tau^2$  controls how similar a dyad's latent positions are between timesteps; and the emission variances  $\sigma_k^2$  controls how similar the frame distributions are for two contexts with identical latent states.

All these prior parameters are learned, thus naturally leading the model to learn highly likely levels of sparsity and variability. This is tremendously convenient in practice, since there are no hyperparameters that need to be tuned (beyond  $K$  and data preprocessing decisions). It also helps the model learn better solutions; for example, Asuncion et al. (2009) finds that Dirichlet concentration learning gives much better solutions for LDA.

The symmetric Dirichlet parameter  $b$  is learned with slice sampling (Neal, 2003), under an improper uniform prior for  $b$ . (In other experiments we have found different diffuse priors for  $b$  make little difference.) Slice sampling only requires an (unnormalized) posterior density function; with a uniform prior it's just the Dirichlet-multinomial likelihood, which is, integrating out  $\phi$ ,

$$L(b) = p(w | z, b) = \prod_{k=1}^K \frac{\Gamma(b)}{\Gamma(b+n_k)} \prod_{w=1}^V \frac{\Gamma(b/V + n_{k,w})}{\Gamma(b/V)} \quad (6.14)$$

where  $V$  is the verb-path vocabulary size,  $n_k$  is the number of event tuples having  $z = k$ , and  $n_{k,w}$  the number having frame  $k$  and verb-path  $w$ . An implementation speedup is possible noting that

---

<sup>15</sup>e.g. <http://www.cs.cmu.edu/~ggordon/10725-F12/slides/11-matrix-newton-annotated.pdf>

<sup>16</sup>In fact, this is not even the factored marginals of  $N(\hat{\eta}, H^{-1})$ , since the diagonal of a Hessian inverse is different than the inverse of a Hessian diagonal.

each frame's lexical count vector ( $n_{k,1}..n_{k,V}$ ) is usually very sparse with mostly 0's, so those terms can be skipped in the innermost loop. (Also draw out the  $\Gamma(b/V)$  denominator.) This sparsity during Gibbs sampling is a natural consequence of the sparsity of language; it can be exploited in other ways to improve sampling efficiency, e.g. Yao et al. (2009).

The context model's variance terms are also learned. We use a conjugate inverse-Wishart prior, in inverse chi-squared parameterization (e.g. Murphy, 2012 §4.6.2.2) of  $\chi^{-2}$ (prior strength, prior value), using diffuse prior  $\chi^{-2}(1, 1)$ . However, since the amount of data is very high, the posterior intervals are very small (often less than  $10^{-3}$ ), and sampling is nearly equivalent to ML inference.

Technically, the conjugate sampling equations are

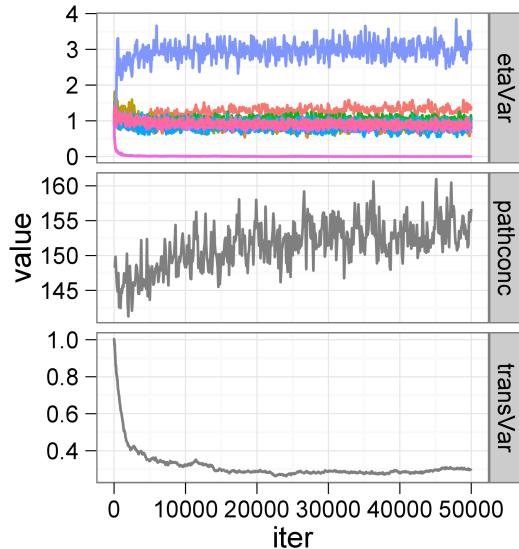
$$\tau^2 \sim \chi^{-2} \left( 1 + N, \frac{1}{1 + N} \left[ 1 + \sum_{s,r,t > 1,k} (\beta_{s,r,t,k} - \beta_{s,r,t-1,k})^2 \right] \right)$$

where  $N$  is  $(K - 1) \times \text{NumDyads} \times (\text{NumTimesteps}-1)$ , and

$$\sigma_k^2 \sim \chi^{-2} \left( 1 + N, \frac{1}{1 + N} \left[ 1 + \sum_{(s,r,t) \text{ where } n_{s,r,t} > 0} (\eta_{s,r,t,k} - \bar{\eta}_{s,r,t,k})^2 \right] \right)$$

where  $N$  is the number of contexts with non-zero events. In both cases  $N$  is hundreds of thousands to millions, swamping the prior pseudocount value of 1.

Here is a plot of the dispersion parameters over one Gibbs sampling run (all 10  $\sigma_k^2$ 's,  $b$ , then  $\tau^2$ ). The fact that dispersions are still drifting in early iterations is an indicator the sampler has not mixed. (Indeed, even though we attained useful results at iteration 10,000, and changing the number of iterations to higher numbers made little difference to the evaluation metrics, these plots clearly indicate mixing has not been achieved at that point. The inferences can be justified only as approximations (but useful ones) of the posterior.)



### 6.10.5 Softmax bug

The results in §6.6.3 have an anomaly where one frame often has a very low probability mass, so it is essentially has only  $K - 1$  latent classes. (This is why the  $K = 2$  models essentially learn

only one class, and thus have a similar lexical scale purity as the random choice baseline that would come from one big cluster of all words.) This was later discovered to be due to a bug in the implementation: it clamped the  $K$ 'th element of  $\eta$  to 0 (attempting to implement an alternate version of softmax with better identifiability), but only the first  $K - 1$  elements of  $\theta$  were used in the posterior density evaluation for the MH step, so counts of  $z = K$  were ignored in the likelihood. Thus the model would eventually shift  $\theta_K$  to zero and put all the probability mass on the first  $K - 1$  elements. (It takes a while for the bug to cause this to happen, since the MAP optimum and Laplace approximation for  $\eta$ , given a fixed  $\theta$ , is computed correctly. But the density ratio for the MH step prefers assigning low  $\theta_K$  values, and in the limit the Markov chain should give zero probability for non-zero  $\theta_K$ .) If  $\theta_K = 0$ , then the model is exactly the same as a fully parameterized  $K - 1$  model; since it is close to zero, it is very similar to that.

## *Chapter 7*

---

# Conclusion

---

## 7.1 Summary of contributions

We have presented a series of tools and case studies that analyze social phenomena and text: examining whether text analysis can help measure or predict social variables, and exploring hypotheses about how social factors guide the text generation process. These hold promise to shed light on a variety of areas of human society where the content of ideas, opinions, events, and other linguistically expressible concepts are crucial for how people behave. As textual and other digital records of human behavior keep growing at a rapid pace, these types of techniques will be key parts of the emerging discipline of computational social science (Lazer et al., 2009).

Specific contributions of this work include:

- The *MiTextExplorer* interactive data analysis system, for exploring the statistical relationship between document text and covariates (Chapter 2). It uses pointwise mutual information, a basic and useful technique for identifying phenomena worthy of further exploration. This is useful for many social text data analysis problems, which typically see document covariates as indicators of social attributes. (The other chapters of this thesis use these and other methods to get at these relationships.) An experimental prototype of this system is open source and available at <http://brenocon.com/mte/>.
- An analysis of public opinion polls compared to opinion expressed in social media, on several economic and political topics (Chapter 3). Some moderate correlations over time are observed, indicating the potential of sentiment analysis as an alternative to polling, but many challenges also exist, and are discussed.
- A mixed-membership model of geography, demographics, and lexical variation (Chapter 4), which infers geographically coherent linguistic communities, characterized by both spatial and word clusters. Applied to geotagged messages in social media, it reveals surprising patterns of geographically specific terms. The model also allows inference of a user's location based on the text of their messages. This model and dataset are also applied to learning demographically-based word groupings, by using U.S. Census data. The exploratory findings here motivate the next chapter:
- A study of linguistic diffusion (Chapter 5), which analyzes how novel terms spread across different cities in the U.S. over several years, as evidenced in social media. Our model for this statistical analysis is robust to confounds in data sampling rates and ephemeral trends.

We find that geography and population size are important drivers of diffusion, but also, demographic similarity—especially with regards to race—is a crucial determinant.

- We present a method to extract events in international relations from a corpus of news articles (Chapter 6), which unlike previous work automatically infers the latent classes of events, as well as their variation across time for different pairs of actors. The models’ inferences are evaluated on their ability to reconstruct previous databases of international conflict, as well as pre-existing phrase dictionaries that have been engineered for this information extraction task. This illustrates the usefulness of natural language processing techniques such as syntactic parsing and the use of semantic arguments in probabilistic modeling.

## 7.2 Recurring themes

Some recurring themes in this work include the following.

**Exploratory analysis.** Many approaches are useful for exploring these types of social text data in light of different questions. Chapter 2 focuses on computationally cheap and fast count-based techniques, which are fast enough to allow rapid iteration and interactive refinement of hypotheses about the relationship of text and document covariates. Chapter 3 also uses count-based techniques for analyzing textual sentiment, but based on pre-existing linguistic dictionaries. In contrast, Chapter 4 develops a more expensive topic model approach, which infers groupings of terms, and Chapter 6 applies similar statistical models to a very different linguistic structure, syntactic dependencies describing events between actors. In all cases, these models allow exploration of social variables against text variables; there is a tradeoff between richness of representation and speed and simplicity of analysis.

**Social science.** An understanding of social science is crucial for useful social text data analysis. When analyzing Twitter data against time and space, many insights from the areas of sociolinguistics and dialectology are important to help interpret what is going on, and computational data analysis can shed light on important hypotheses in this area (Chapters 4 and 5). Sentiment analysis on Twitter is useful to think about in comparison to traditional polling methods (Chapter 3); there are of course many differences in how these measurement methods work. In some cases, previous work additionally provides rich datasets to work with; for example, the area of international relations has seen much work that we use to help validate and understand what our new event extraction methods methods are doing (Chapter 6), and in future work they may be integrated in new approaches to yield better analysis techniques.

**Probabilistic machine learning.** The framework of probabilistic graphical models and statistical modeling is used throughout. This gives a rich set of formal and computational tools to represent, infer, and learn important relationships in the data. Even when very simple techniques are used, it is enlightening to keep in mind statistical and machine learning interpretations of them—as in Chapter 2’s use of pointwise mutual information for word analysis. Optimization, variational inference, and Markov chain Monte Carlo methods allow for flexible and useful inferences for a very wide variety of problems. A few other important techniques used in this thesis include false discovery rate control, dynamic programming, and distributed computation. These all have important uses in the social sciences.

**A little bit of NLP can go a long way.** Techniques for text analysis (i.e., natural language processing) are quite imperfect given the incredible complexity of language and human intelligence, but they can extract partial views of sentiment, entities, and events, which have many important applications in understanding social phenomena revealed in text data.

### 7.3 Directions for future work

Much of this work has a heavily exploratory nature. This is appropriate for early stages of gaining understanding, especially as we are still coming to grips with the remarkable possibilities of new social data produced by online systems.

However, the real gains will be made through deeper integration with substantive issues in the social sciences. One relatively simple approach is to leverage data and resources from previous social science research to guide the development of new algorithms and techniques. Chapter 3 explores to what extent traditional polling data might correlate to sentiment measured in social media, and Chapter 6 seeks to reconstruct traditional international event data from latent syntactic analysis techniques. This type of research has the attraction of a relatively simple framework—the researcher checks whether a new computational method correlates to an older, established method. In the longer term, it allows a relatively loose interaction between new computational research and more substantive research efforts: the computational researchers run algorithms to generate new data, then hand it off to others for further analysis. Many examples of this pattern exist: previous work in international event data, for example, follows this approach (King and Lowe, 2003; Schrodt, 2012), and a future version of our international events model could take a similar path. In political science, the well-known NOMINATE scores, a latent variable model of legislator partisanship (Poole and Rosenthal, 1997), are similarly inferred by a small set of researchers, and then made available for download; and they have indeed been used very widely.<sup>1</sup>

A more involved, in-depth integration of computation and social science is to directly engage substantive questions as part of the computational analysis. One avenue is by developing general-purpose text analysis tools that are directly used by researchers—methods such as keyword queries, dictionary frequencies, document classifiers, and topic models have all been widely used—and new methods such as the interactive explorer of Chapter 2 may help as well. Here, the computational research has substantive researchers as users; an iterative process of feedback and refinement is crucial for future progress.

Another approach is to develop customized, computationally-heavy analysis for a particular question, such as a statistical model of text conditioned on certain types of social variables (as in the generative process described in §1.1). In the course of our research on sociolinguistics and Twitter, we found that a latent variable, temporal model was necessary to investigate questions about linguistic diffusion (Chapter 5). To support this form of development, contributions from the computational, statistical, and linguistic sciences include the fundamental primitives of model building (such as probabilistic graphical models), and defining the linguistic objects of analysis. General advances in these areas, such as automatic statistical modeling frameworks like BUGS (Lunn et al., 2009) or Stan (Stan Development Team, 2014), or more robust natural language processing tools, can make the process of model development easier.

An exciting frontier is to incorporate ideas of measurement, causal inference, and other core social science topics into text analysis models. As mentioned in §3.6, the process by which opinions become expressed in social media is subject to a large number of complicated behavioral factors. In order to formulate more effective opinion measurement methods, a promising approach is to

---

<sup>1</sup><http://www.voteview.com/>

model the data generating process, where latent variables of user and population opinions drive a process resulting in frequencies of different types of messages in social media. For international events data generated by traditional rule-based systems, Lowe (2012) examines the problem of statistical aggregation by positing a latent variable measurement model, where event counts are generated based on the dyadic relationship; from this a scaling of event classes is inferred. Our event class model may benefit from an approach like this, or other alternative models of international actors and their dynamics. These can be combined with sophisticated linguistic representations. Researchers have also started to examine causal inference with text data (Roberts et al., 2014); this is a critical area for future work. Indeed, as many interesting “big data” possibilities emerge from digital records of human behavior from sensors or online environments, we will see more and more need for social analysis; as Grimmer (2014) puts it, “we’re all social scientists now.”

In terms of methodology, it might be useful to think of the following goal: make text analysis as widespread and useful as linear regression, the workhorse of much applied statistical analysis. Though it is not always thought of in this way, linear regression is a completely computational methodology—before computers, it was impossible to fit the types of linear regressions that are now commonplace. Since linear regressions and related statistical analysis methods are well understood, and widely implemented and straightforward to run, their use is usually not considered “computational social science” in the way that, say, topic models or network analyses sometimes are. But this is only because the methodology of linear regression is more mature than many text analysis models.

Finally, this line of research should not just provide better social science methodology, but feed back to natural language and artificial intelligence research as well. Language and cognition are deeply embedded in social processes: communities and their shared norms and ideas are key to giving language meaning. The social context for an utterance is an important piece of knowledge that should be integrated into models of language production and understanding. Research into analysis of social text data will reveal important considerations and insights for computational intelligence more generally.

---

# Bibliography

---

Cell Phone Task Force AAPOR. New considerations for survey researchers when planning and conducting RDD telephone surveys in the US with respondents reached via cell phone numbers. American Association for Public Opinion Research, 2010. URL [http://www.aapor.org/Cell\\_Phone\\_Task\\_Force\\_Report.htm](http://www.aapor.org/Cell_Phone_Task_Force_Report.htm).

Amr Ahmed, Liangjie Hong, and Alexander J. Smola. Hierarchical geographical modeling of user locations from social media posts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 25–36. International World Wide Web Conferences Steering Committee, 2013.

Eduardo G. Altmann, Janet B. Pierrehumbert, and Adilson E. Motter. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLoS ONE*, 4(11):e7678+, November 2009. doi: 10.1371/journal.pone.0007678. URL <http://dx.doi.org/10.1371/journal.pone.0007678>.

Anima Anandkumar, Yi-kai Liu, Daniel J. Hsu, Dean P. Foster, and Sham M. Kakade. A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 917–925, 2012.

Jannis K. Androutsopoulos. Non-standard spellings in media texts: The case of German fanzines. *Journal of Sociolinguistics*, 4(4):514–533, 2000.

Jacques Anis. Neography: Unconventional spelling in French SMS text messages. In Brenda Danet and Susan C. Herring, editors, *The Multilingual Internet: Language, Culture, and Communication Online*, pages 87 – 115. Oxford University Press, 2007.

Anonymous. The junk science behind the ‘Twitter Hedge Fund’, 2012. URL <http://sellthenews.tumblr.com/post/21067996377/noitdoesnot>.

Francis J. Anscombe. Graphs in statistical analysis. *The American Statistician*, 27(1):17–21, 1973.

Werner Antweiler and Murray Z. Frank. Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, 59(3):1259–1294, June 2004. ISSN 00221082. URL <http://www.jstor.org/stable/3694736>.

Sinan Aral and Dylan Walker. Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341, July 2012. ISSN 1095-9203. doi: 10.1126/science.1215842. URL <http://dx.doi.org/10.1126/science.1215842>.

Nikolay Archak, Anindya Ghose, and Panagiotis Ipeirotis. Deriving the pricing power of product features by mining consumer reviews. *Management Science*, page mnsc1110, 2011.

Shlomo Argamon, Charles Cooney, Russell Horton, Mark Olsen, Sterling Stein, and Robert Voyer. Gender, race, and nationality in black drama, 1950-2006: Mining differences in language use in authors and their characters. *Digital Humanities Quarterly*, 3(2), 2009.

Sanjeev Arora, Rong Ge, Yoni Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. *arXiv preprint arXiv:1212.4777*, 2012.

Nikolaos Askitas and Klaus F. Zimmermann. Google econometrics and unemployment forecasting. *Applied Economics Quarterly*, 55(2):107–120, April 2009. ISSN 1611-6607. doi: 10.3790/aeq.55.2.107. URL <http://www.atypon-link.com/DH/doi/abs/10.3790/aeq.55.2.107>.

Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. On smoothing and inference for topic models. In *Uncertainty in Artificial Intelligence*, volume 100, 2009.

Sitaram Asur and Bernardo A. Huberman. Predicting the future with social media. *arXiv:1003.5699*, March 2010. URL <http://arxiv.org/abs/1003.5699>.

Edward E. Azar and Thomas Sloan. Dimensions of interactions. Technical report, University Center of International Studies, University of Pittsburgh, Pittsburgh, 1975.

L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak. Spatial variation in search engine queries. In *Proceedings of the International World Wide Web Conference (WWW)*, 2008.

Charles-James Bailey. *Variation and Linguistic Theory*. Center for Applied Linguistics, 1973.

Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 519–528, 2012.

David Bamman and Gregory Crane. Measuring historical word sense variation. In *Proceeding of the 11th annual international ACM/IEEE joint conference on Digital libraries*, page 110, 2011.

David Bamman, Brendan O’Connor, and Noah A. Smith. Censorship and deletion practices in Chinese social media. *First Monday*, 17(3), 2012.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160, 2014.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. Open Information Extraction from the Web. *IJCAI*, 2007.

G. J. Baxter, R. A. Blythe, W. Croft, and A. J. McKane. Utterance selection model of language change. *Physical Review E*, 73:046118+, April 2006. doi: 10.1103/PhysRevE.73.046118. URL <http://dx.doi.org/10.1103/PhysRevE.73.046118>.

Nick Beauchamp. Predicting and interpolating state-level polling using Twitter textual data. New Directions in Analyzing Text as Data Workshop, 2013. URL [http://www.kenbenoit.net/pdfs/NDATAD2013/Beauchamp\\_twitterpolls\\_2.pdf](http://www.kenbenoit.net/pdfs/NDATAD2013/Beauchamp_twitterpolls_2.pdf).

Richard A. Becker and John M. Chambers. *S: an interactive environment for data analysis and graphics*. CRC Press, 1984.

Richard A. Becker and William S. Cleveland. Brushing scatterplots. *Technometrics*, 29(2):127–142, 1987.

Cosmin Adrian Bejan. Unsupervised discovery of event scenarios from texts. In *Proceedings of the 21st Florida Artificial Intelligence Research Society International Conference (FLAIRS)*, Coconut Grove, FL, USA, 2008.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

Steven Bethard and Dan Jurafsky. Who should I cite: learning literature search models from citation behavior. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, page 609618, 2010.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Ryan C. Black, Sarah A. Treul, Timothy R. Johnson, and Jerry Goldman. Emotions, oral arguments, and Supreme Court decision making. *The Journal of Politics*, 73(2):572–581, April 2011.

David Blei and Jon McAuliffe. Supervised topic models. In *Neural Information Processing Systems*, pages 121–128, Cambridge, MA, 2008. MIT Press.

David M. Blei and Michael I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1:121–144, 2006.

David M. Blei and John D. Lafferty. Correlated topic models. In *Neural Information Processing Systems*, 2006a.

David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of ICML*, 2006b.

David M. Blei and John D. Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, 1(1):17–35, 2007.

David M. Blei and John D. Lafferty. Topic models. *Text mining: classification, clustering, and applications*, 10:71, 2009.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250, Vancouver, Canada, 2008. ACM. ISBN 978-1-60558-102-6. URL <http://dx.doi.org/10.1145/1376616.1376746>.

Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. 1010.3003, October 2010. URL <http://arxiv.org/abs/1010.3003>.

Robert M. Bond, Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298, September 2012. ISSN 0028-0836. doi: 10.1038/nature11421. URL <http://dx.doi.org/10.1038/nature11421>.

Elizabeth Boschee, Premkumar Natarajan, and Ralph Weischedel. Automatic extraction of events from open source text for predictive forecasting. *Handbook of Computational Approaches to Counterterrorism*, page 51, 2013.

Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D3: Data-driven documents. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2011. URL <http://vis.stanford.edu/papers/d3>.

Remco Bouckaert, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson. Mapping the origins and expansion of the indo-european language family. *Science*, 337(6097):957–960, 2012.

Patrick T. Brandt, John R. Freeman, and Philip A. Schrodt. Real time, time series forecasting of inter-and intra-state political conflict. *Conflict Management and Peace Science*, 28(1):41–64, 2011.

Patrick T. Brandt, John R. Freeman, Tse-min Lin, and Phillip A. Schrodt. A Bayesian time series approach to the comparison of conflict dynamics. In *APSA 2012 Annual Meeting Paper*, 2012.

David A. Broniatowski, Michael J. Paul, and Mark Dredze. National and local influenza surveillance through twitter: An analysis of the 2012-2013 influenza epidemic. *PloS one*, 8(12):e83672, 2013.

Mary Bucholtz, Nancy Bermudez, Victor Fung, Lisa Edwards, and Rosalva Vargas. Hella Nor Cal or totally So Cal? the perceptual dialectology of California. *Journal of English Linguistics*, 35(4): 325–352, 2007. URL <http://people.duke.edu/~eec10/hellanorcal.pdf>.

Andreas Buja, John Alan McDonald, John Michalak, and Werner Stuetzle. Interactive data visualization using focusing and linking. In *Visualization, 1991. Visualization'91, Proceedings., IEEE Conference on*, pages 156–163. IEEE, 1991.

Andreas Buja, Dianne Cook, and Deborah F. Swayne. Interactive high-dimensional data visualization. *Journal of Computational and Graphical Statistics*, 5(1):78–99, 1996.

John Burrows. All the way through: testing for authorship in different frequency strata. *Literary and Linguistic Computing*, 22(1):27–47, 2007.

Olivier Cappe, Simon J. Godsill, and Eric Moulines. An overview of existing methods and recent advances in sequential Monte Carlo. *Proceedings of the IEEE*, 95(5):899–924, May 2007. ISSN 0018-9219. doi: 10.1109/JPROC.2007.893250. URL <http://dx.doi.org/10.1109/JPROC.2007.893250>.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevan R. Hruschka, and Tom M. Mitchell. Toward an architecture for never-ending language learning. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, pages 1306–1313, 2010.

Chris K. Carter and Robert Kohn. On Gibbs sampling for state space models. *Biometrika*, 81(3): 541–553, 1994.

F. G. Cassidy and J. H. Hall. *Dictionary of American Regional English*, volume 1. Harvard University Press, 1985.

Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P. Krishna Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the International Workshop on Web and Social Media (ICWSM)*, pages 10–17, 2010.

J. Chambers. *Sociolinguistic Theory: Linguistic Variation and its Social Significance*. Blackwell, 2009.

Nathanael Chambers and Dan Jurafsky. Unsupervised learning of narrative schemas and their participants. In *Proceedings of ACL-IJCNLP*. Association for Computational Linguistics, 2009.

Nathanael Chambers and Dan Jurafsky. Template-based information extraction without the templates. In *Proceedings of ACL*, 2011.

Allison J.B. Chaney and David M. Blei. Visualizing topic models. In *Proceedings of ICWSM*, 2013.

Jonathan Chang, Jordan Boyd-Graber, and David M. Blei. Connections between the lines: augmenting social networks with text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 169–178. ACM, 2009.

Jonathan Chang, Itamar Rosenn, Lars Backstrom, and Cameron Marlow. epluribus: Ethnicity on social networks. In *Proceedings of the International Workshop on Web and Social Media (ICWSM)*, volume 10, pages 18–25, 2010.

L. C. Chang and J. A. Krosnick. National surveys via RDD telephone interviewing vs. the internet: Comparing sample representativeness and response quality. *Manuscript under review*, 2003.

Jackie Chi Kit Cheung, Hoifung Poon, and Lucy Vanderwende. Probabilistic frame induction. In *Proceedings of NAACL*, 2013. arXiv preprint arXiv:1302.4813.

Jason Chuang, Christopher D. Manning, and Jeffrey Heer. Termite: Visualization techniques for assessing textual topic models. In *Advanced Visual Interfaces*, 2012. URL <http://vis.stanford.edu/papers/termite>.

K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):2229, 1990.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

William S. Cleveland. *Visualizing data*. Hobart Press, 1993.

Dan Cohen, Frederick Gibbs, Tim Hitchcock, Geoffrey Rockwell, Jorg Sander, Robert Shoemaker, Stefan Sinclair, Sean Takats, William J. Turkel, Cyril Briquet, Jamie McLaughlin, Milena Radzikowska, John Simpson, and Kirsten C. Uszkalo. Data mining with criminal intent. White paper, 2011. URL <http://criminalintent.org>.

Dianne Cook and Deborah F. Swayne. *Interactive and dynamic graphics for data analysis: with R and GGobi*. Springer, 2007.

Samantha R. Cook, Andrew Gelman, and Donald B. Rubin. Validation of software for bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, 15(3), 2006.

- D. Hugh Craig and Arthur F. Kinney. *Shakespeare, computers, and the mystery of authorship*. Cambridge University Press, 2009.
- D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *Proceedings of the International World Wide Web Conference (WWW)*, 2009.
- David Crystal. *Language and the Internet*. Cambridge University Press, second edition, September 2006.
- Aron Culotta. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the first workshop on social media analytics*, pages 115–122. ACM, 2010.
- Sanjiv R. Das and Mike Y. Chen. Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388, September 2007. doi: 10.1287/mnsc.1070.0704. URL <http://mansci.journal.informs.org/cgi/content/abstract/53/9/1375>.
- Hal Daumé III. Non-parametric Bayesian areal linguistics. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 593–601. Association for Computational Linguistics, 2009.
- M. C. de Marneffe and C. D. Manning. The stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, page 18, 2008.
- Jana Diesner and Kathleen M. Carley. Revealing social structure from texts: meta-matrix text analysis as a novel method for network text analysis. In *Causal mapping for information systems and technology research*, pages 81–108. Harrisburg, PA: Idea Group Publishing, 2005.
- Peter S. Dodds and Christopher M. Danforth. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, page 116, 2009.
- Peter Sheridan Dodds, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS ONE*, 6(12):e26752, 2011.
- Michael R. Dressman. Redd up. *American Speech*, 54(2):141–145, 1979.
- Maeve Duggan and Aaron Smith. Social media update 2013. Technical report, Pew Research Center, December 2013.
- Michael Dunn, Simon J. Greenhill, Stephen C. Levinson, and Russell D. Gray. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345):79–82, 2011.
- Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19:61–74, 1993. doi: 10.1.1.14.5962.
- Penelope Eckert. *Jocks and Burnouts: Social Categories and Identity in the High School*. Teachers College Press, 1989.
- J. Eisenstein, A. Ahmed, and E.P. Xing. Sparse additive generative models of text. In *Proceedings of ICML*, pages 1041–1048, 2011a.
- Jacob Eisenstein. Geographic topic model: Appendix, 2010. URL [http://www.cc.gatech.edu/~jeisenst/papers/emnlp2010\\_appendix.pdf](http://www.cc.gatech.edu/~jeisenst/papers/emnlp2010_appendix.pdf).

Jacob Eisenstein, Brendan O’Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, 2010.

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. Sparse additive generative models of text. In *Proceedings of ICML*, pages 1041–1048, 2011b.

Jacob Eisenstein, Noah A. Smith, and Eric P. Xing. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1365–1374. Association for Computational Linguistics, 2011c.

Jacob Eisenstein, Brendan O’Connor, Noah A. Smith, and Eric P. Xing. Mapping the geographical diffusion of new words. In *NIPS Workshop on Social Network and Social Media Analysis*, 2012. URL <http://arxiv.org/abs/1210.5268>.

E. Erosheva, S. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5220, 2004.

Zsuzsanna Fagyal, Samarth Swarup, Anna M. Escobar, Les Gasser, and Kiran Lakkaraju. Centers and peripheries: Network roles in language change. *Lingua*, 120(8):2061–2079, August 2010. ISSN 00243841. doi: 10.1016/j.lingua.2010.02.001. URL <http://dx.doi.org/10.1016/j.lingua.2010.02.001>.

Carol Friedman, Pauline Kra, Hong Yu, Michael Krauthammer, and Andrey Rzhetsky. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17(suppl 1):S74–S82, 2001.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), January 2010.

M.J. Gardner, J. Lutes, J. Lund, J. Hansen, D. Walker, E. Ringger, and K. Seppi. The topic browser: An interactive tool for browsing topic models. In *NIPS Workshop on Challenges of Data Visualization*. MIT Press, 2010.

Daniel Gayo-Avello. I wanted to predict elections with Twitter and all i got was this lousy paper: a balanced survey on election prediction using Twitter data. *arXiv preprint arXiv:1204.6441*, 2012.

Arthur Gelb. *Applied Optimal Estimation*. MIT press, 1974.

Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 1st edition, 2006. ISBN 052168689X.

Andrew Gelman and Cosma Rohilla Shalizi. Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38, 2013.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian data analysis*. Chapman & Hall/CRC, 2003.

Matthew Gentzkow and Jesse M. Shapiro. What drives media slant? Evidence from U.S. daily newspapers. *Econometrica*, 78(1):35–71, 2010.

Deborah J. Gerner, Philip A. Schrodt, Omur Yilmaz, and Rajaa Abu-Jabr. The Creation of CAMEO (Conflict and Mediation Event Observations): An Event Data Framework for a Post Cold War World. *Annual Meeting of the American Political Science Association*, 2002.

Sean M. Gerrish. *Applications of Latent Variable Models in Modeling Influence and Decision Making*. PhD thesis, Princeton University, 2013.

Sean M. Gerrish and David M. Blei. A language-based approach to measuring scholarly impact. In *Proceedings of ICML Workshop on Computational Social Science*, 2010.

Sean M. Gerrish and David M. Blei. Predicting legislative roll calls from text. In *Proceedings of ICML*, 2011.

Faten Ghosn, Glenn Palmer, and Stuart A. Bremer. The MID3 data set, 1993–2001: Procedures, coding rules, and description. *Conflict Management and Peace Science*, 21(2):133–154, 2004.

Eric Gilbert and Karrie Karahalios. Widespread worry and the stock market. In *Proceedings of the International Conference on Weblogs and Social Media*, 2010.

Daniel Gildea. Probabilistic models of verb-argument structure. In *Proceedings of COLING*, 2002.

Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, February 2009. ISSN 0028-0836. doi: 10.1038/nature07634. URL <http://dx.doi.org/10.1038/nature07634>.

Simon J. Godsill, Arnaud Doucet, and Mike West. Monte Carlo smoothing for non-linear time series. In *Journal of the American Statistical Association*, pages 156–168, 2004.

Scott A. Golder and Michael W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333:1878–1881, September 2011. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1202775. URL <http://www.sciencemag.org/cgi/doi/10.1126/science.1202775>.

Joshua S. Goldstein. A conflict-cooperation scale for WEIS events data. *Journal of Conflict Resolution*, 36:369–385, 1992.

Joshua S. Goldstein, Jon C. Pevehouse, Deborah J. Gerner, and Shibley Telhami. Reciprocity, triangularity, and cooperation in the middle east, 1979-97. *Journal of Conflict Resolution*, 45(5):594–620, 2001.

Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. *ACM Transactions on Knowledge Discovery from Data*, 5(4):21, 2012.

Bruno Gonçalves, Nicola Perra, and Alessandro Vespignani. Modeling users' activity on twitter networks: Validation of dunbar's number. *PloS one*, 6(8):e22656, 2011.

Matthew J. Gordon. Phonological correlates of ethnic identity: Evidence of divergence? *American Speech*, 75(2):115–136, 2000.

Carsten Görg, Zhicheng Liu, and John Stasko. Reflections on the evolution of the Jigsaw visual analytics system. *Information Visualization*, 2013.

- Russell D. Gray and Quentin D. Atkinson. Language-tree divergence times support the anatolian theory of indo-european origin. *Nature*, 426(6965):435–439, 2003.
- Russell D. Gray, Alexei J. Drummond, and Simon J. Greenhill. Language phylogenies reveal expansion pulses and pauses in pacific settlement. *science*, 323(5913):479–483, 2009.
- T. Grenager and C. D. Manning. Unsupervised discovery of a statistical verb lexicon. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, page 18, 2006. ISBN 1932432736.
- T.L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228, 2004.
- Justin Grimmer. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1, 2010.
- Justin Grimmer. We’re all social scientists now: How big data, machine learning, and causal inference work together, 2014. URL [http://stanford.edu/~jgrimmer/bd\\_2.pdf](http://stanford.edu/~jgrimmer/bd_2.pdf).
- Justin Grimmer and Gary King. General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences*, 108(7):2643–2650, 2011.
- Justin Grimmer and Brandon M. Stewart. Text as Data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297, 2013. URL <http://www.stanford.edu/~jgrimmer/tad2.pdf>.
- Robert L. Grossman, Matthew Greenway, Allison P. Heath, Ray Powell, Rafael D. Suarez, Walt Wells, Kevin P. White, Malcolm P. Atkinson, Iraklis A. Klampanos, Heidi L. Alvarez, Christine Harvey, and Joe Mambretti. The design of a community science cloud: The open science data cloud perspective. In *SC Companion*, pages 1051–1057, 2012.
- Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.
- Jeff Harrison and Mike West. *Bayesian forecasting and dynamic models*. Springer Verlag, New York, 1997.
- Marti Hearst. *Search user interfaces*. Cambridge University Press, 2009.
- Brent Hecht and Monica Stephens. A tale of cities: Urban biases in volunteered geographic information. In *Proceedings of the International Workshop on Web and Social Media (ICWSM)*, 2014.
- Susan C. Herring. Grammar and electronic communication. In Carol A. Chapelle, editor, *The Encyclopedia of Applied Linguistics*. Wiley, 2012.
- Jerry R. Hobbs, Douglas Appelt, John Bear, David Israel, Megumi Kameyama, Mark Stickel, and Mabry Tyson. Fastus: A cascaded finite-state transducer for extracting information from natural-language text. *Finite-State Language Processing*, page 383, 1997.
- Peter D. Hoff. Nonparametric modeling of hierarchically exchangeable data. *University of Washington Statistics Department, Technical Report*, 421, 2003.
- Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Chris C. Holmes and Leonhard Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1):145–168, 2006.

David I. Holmes. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3):111–117, 1998. doi: 10.1093/linc/13.3.111. URL <http://llc.oxfordjournals.org/content/13/3/111.abstract>.

Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsoutsouliklis. Discovering geographical topics in the Twitter stream. In *Proceedings of the 21st international conference on World Wide Web*, pages 769–778. ACM, 2012.

Daniel J. Hopkins and Gary King. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247, January 2010. ISSN 00925853. doi: 10.1111/j.1540-5907.2009.00428.x. URL <http://dash.harvard.edu/handle/1/4142694>.

Russell Horton, Robert Morrissey, Mark Olsen, Glenn Roe, and Robert Voyer. Mining Eighteenth Century Ontologies: Machine Learning and Knowledge Classification in the Encyclopédie. *Digital Humanities Quarterly*, 3(2), 2009. URL <http://www.digitalhumanities.org/dhq/vol/3/2/000044.html>.

Mark Edward Huberty. Multi-cycle forecasting of congressional elections with social media. In *Proceedings of the 2nd workshop on Politics, Elections and Data*, pages 23–30. ACM, 2013.

Matthew L. Jockers. *Macroanalysis: Digital methods and literary history*. University of Illinois Press, 2013.

D. E. Johnson. Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass*, 3(1):359–383, 2009. ISSN 1749818X. doi: 10.1111/j.1749-818X.2008.00108.x. URL <http://blackwell-synergy.com/doi/abs/10.1111/j.1749-818X.2008.00108.x>.

B. Johnstone. Language and place. In R. Mesthrie and W. Wolfram, editors, *Cambridge Handbook of Sociolinguistics*. Cambridge University Press, 2010.

D.M. Jones, S.A. Bremer, and J.D. Singer. Militarized interstate disputes, 1816–1992: Rationale, coding rules, and empirical patterns. *Conflict Management and Peace Science*, 15(2):163–213, 1996.

Mahesh Joshi, Dipanjan Das, Kevin Gimpel, and Noah A. Smith. Movie reviews and revenues: An experiment in text regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, page 293296, 2010.

Andreas Jungherr, Pascal Jürgens, and Harald Schoen. Why the Pirate Party won the German election of 2009 or the trouble with predictions: A response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. “predicting elections with twitter: What 140 characters reveal about political sentiment”. *Social Science Computer Review*, 30(2):229–234, 2012.

Matthew E. Kahn and Matthew J. Kotchen. Environmental concern and the business cycle: The chilling effect of recession. <http://www.nber.org/papers/w16241>, July 2010. URL <http://www.nber.org/papers/w16241>.

Adam Kilgarriff. Comparing corpora. *International journal of corpus linguistics*, 6(1):97–133, 2001.

Gary King and Will Lowe. An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization*, 57(3):617–642, July 2003.

Gary King, Robert O. Keohane, and Sidney Verba. *Designing social inquiry: Scientific inference in qualitative research*. Princeton University Press, 1994.

Gary King, Jennifer Pan, and Margaret E. Roberts. How censorship in china allows government criticism but silences collective expression. *American Political Science Review*, 107:1–18, 2013.

Shimon Kogan, Dmitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, page 272280, 2009.

M. Koppel and I. Shtrimberg. Good news or bad news? let the market decide. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, 2004.

Klaus Krippendorff. *Content analysis: an introduction to its methodology*. SAGE Publications, Inc., 2012.

J. A. Krosnick, C. M. Judd, and B. Wittenbrink. The measurement of attitudes. In *The Handbook of Attitudes*, page 2176. Psychology Press, 2005.

Hans Kurath. *A Word Geography of the Eastern United States*. University of Michigan Press, 1949.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proceedings of the International World Wide Web Conference (WWW)*, pages 591–600, 2010.

William Labov. *The Social Stratification of English in New York City*. Center for Applied Linguistics, 1966.

William Labov, editor. *Locating Language in Time and Space*. Academic Press, 1980.

William Labov. *Principles of Linguistic Change, Volume 2: Social Factors*. Blackwell, 2001.

William Labov. Pursuing the cascade model. In David Britain and Jenny Cheshire, editors, *Social Dialectology: In honour of Peter Trudgill*. John Benjamins, 2003.

William Labov, Sharon Ash, and Charles Boberg. *The Atlas of North American English: Phonetics, Phonology, and Sound Change*. Walter de Gruyter, 2006.

Joel Lang and Mirella Lapata. Unsupervised induction of semantic roles. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 939–947. Association for Computational Linguistics, 2010.

Victor Lavrenko, Matt Schmill, Dawn Lawrie, Paul Ogilvie, David Jensen, and James Allan. Mining of concurrent text and time series. In *Proceedings of KDD Workshop on Text Mining*, pages 37–44, 2000.

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Lszl Barabsi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Computational social science. *Science*, 323(5915):721–723, February 2009. doi: 10.1126/science.1167742. URL <http://www.scienmag.org/content/323/5915/721.short>.

Kyumin Lee, Brian David Eoff, and James Caverlee. Seven months with the devils: A long-term study of content polluters on Twitter. In *Proceedings of the International Workshop on Web and Social Media (ICWSM)*, 2011.

Margaret G. Lee. Out of the hood and into the news: Borrowed black verbal expressions in a mainstream newspaper. *American Speech*, 74(4):369–388, 1999.

Wendy G. Lehnert. Cognition, computers, and car bombs: How Yale prepared me for the 1990s. In *Beliefs, Reasoning, and Decision-Making. Psycho-Logic in Honor of Bob Abelson*, pages 143–173, Hillsdale, NJ, Hove, UK, 1994. Erlbaum. <http://ciir.cs.umass.edu/pubfiles/cognition3.pdf>.

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 497–506, 2009.

H. Li, X. Li, H. Ji, and Y. Marton. Domain-independent novel event discovery and semi-automatic event annotation. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, Sendai, Japan, November*, 2010.

Roddy Lindsay. Predicting polls with lexicon, October 2008. URL <http://web.archive.org/web/20090805103621/http://languagewrong.tumblr.com/post/55722687/predicting-polls-with-lexicon>.

Gilad Lotan, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, and danah boyd. The revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *International Journal of Communication*, 5:1375–1405, 2011.

Tim Loughran and Bill McDonald. When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. *Journal of Finance (forthcoming)*, 2011. URL [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=1331573](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1331573).

Will Lowe. Measurement models for event data. Available at SSRN 2208434, 2012.

Sydney C. Ludvigson. Consumer confidence and consumer spending. *The Journal of Economic Perspectives*, 18(2):29–50, 2004. ISSN 08953309. URL <http://www.jstor.org/stable/3216889>. ArticleType: primary\_article / Full publication date: Spring, 2004 / Copyright 2004 American Economic Association.

David Lunn, David Spiegelhalter, Andrew Thomas, and Nicky Best. The bugs project: Evolution, critique and future directions. *Statistics in medicine*, 28(25):3049–3067, 2009.

Gary Lupyan and Rick Dale. Language structure is partly determined by social structure. *PloS ONE*, 5(1):e8559, 2010.

David J. C. MacKay and Linda C. Baumana Peto. A hierarchical Dirichlet language model. *Natural language engineering*, 1(03):289–308, 1995.

Christopher D. Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT Press, 1999.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 1st edition, July 2008. ISBN 0521865719.

Allen R. Martin and Matthew O. Ward. High dimensional brushing for interactive exploration of multivariate data. In *Proceedings of the 6th Conference on Visualization'95*, page 271. IEEE Computer Society, 1995.

Andrew D. Martin and Kevin M. Quinn. Dynamic ideal point estimation via Markov chain Monte Carlo for the U.S. Supreme Court, 1953–1999. *Political Analysis*, 10(2):134–153, 2002.

C.A. McClelland. Some effects on theory from the international event analysis movement. 1970. Mimeo, University of Southern California.

Q. Mei, C. Liu, H. Su, and C. X. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of the International World Wide Web Conference (WWW)*, 2006.

Q. Mei, X. Ling, M. Wondra, H. Su, and C. X. Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the International World Wide Web Conference (WWW)*, 2007. ISBN 978-1-59593-654-7. doi: <http://doi.acm.org/10.1145/1242572.1242596>.

Marina Meilă. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, 2007.

Sergey Melnik, Andrey Gubarev, Jing Jing Long, Geoffrey Romer, Shiva Shivakumar, Matt Tolton, and Theo Vassilakis. Dremel: interactive analysis of web-scale datasets. *Proceedings of the VLDB Endowment*, 3(1-2):330–339, 2010.

Panagiotis T. Metaxas and Eni Mustafaraj. Social media and the elections. *Science*, 338(6106):472–473, 2012.

Panagiotis T. Metaxas, Eni Mustafaraj, and Daniel Gayo-Avello. How (Not) to predict elections. In *Privacy, security, risk and trust (PASSAT), IEEE Third International Conference on Social Computing (SocialCom)*, 2011.

Lesley Milroy. *Language and Social Networks*. Wiley-Blackwell, 2 edition, 1991.

David Mimno. *Topic regression*. PhD thesis, University of Massachusetts Amherst, 2012.

David Mimno, Hanna Wallach, and Andrew McCallum. Gibbs sampling for logistic normal topic models with graph-based priors. In *NIPS Workshop on Analyzing Graphs*, 2008.

Thomas P. Minka. Estimating a Dirichlet distribution. Technical report, Massachusetts Institute of Technology, 2003. URL <http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/minka-dirichlet.pdf>.

Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. Understanding the demographics of twitter users. In *Proceedings of the International Workshop on Web and Social Media (ICWSM)*, pages 554–557, 2011.

Lewis Mitchell, Morgan R. Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M. Danforth. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE*, 8(5):e64417, 05 2013. doi: 10.1371/journal.pone.0064417. URL <http://dx.doi.org/10.1371%2Fjournal.pone.0064417>.

Ashutosh Modi, Ivan Titov, and Alexandre Klementiev. Unsupervised induction of frame-semantic representations. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, pages 1–7. Association for Computational Linguistics, 2012.

B. L. Monroe, M. P. Colaresi, and K. M. Quinn. Fightin' Words: lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372, 2008.

Fred Morstatter, Jurgen Pfeffer, Huan Liu, and Kathleen M. Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *Proceedings of the International Workshop on Web and Social Media (ICWSM)*, pages 400–408, 2013.

Frederick Mosteller and David Wallace. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, 1964.

Kevin P. Murphy. *Machine Learning: a Probabilistic Perspective*. MIT Press, 2012.

Radford M. Neal. Slice sampling. *Annals of statistics*, pages 705–741, 2003.

John Nerbonne. Data-driven dialectology. *Language and Linguistics Compass*, 3(1):175–198, 2009.

John Nerbonne and Wilbert Heeringa. Geographic distributions of linguistic variation reflect dynamics of differentiation. *Roots: linguistics in search of its evidential base*, 96:267, 2007.

D. Newman, T. Baldwin, L. Cavedon, E. Huang, S. Karimi, D. Martinez, F. Scholer, and J. Zobel. Visualizing search results and document collections using topic maps. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(2):169–175, 2010.

David Newman, Chaitanya Chemudugunta, and Padhraic Smyth. Statistical entity-topic models. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 680–686. ACM, 2006.

Partha Niyogi and Robert C. Berwick. A dynamical systems model for language change. *Complex Systems*, 11(3):161–204, 1997.

Diarmuid Ó Séaghdha. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 435–444. Association for Computational Linguistics, 2010.

Sean P. O'Brien. Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review*, 12(1):87–104, 2010.

Brendan O'Connor. MiTextExplorer: Linked brushing and mutual information for exploratory text data analysis. In *Proceedings of the ACL Workshop on Interactive Language Learning, Visualization, and Interfaces*, 2014.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *International AAAI Conference on Weblogs and Social Media, Washington, DC*, 2010a.

Brendan O'Connor, Jacob Eisenstein, Eric P. Xing, and Noah A. Smith. A mixture model of demographic lexical variation. In *NIPS Workshop on Machine Learning for Social Computing*, 2010b.

Brendan O'Connor, Michel Krieger, and David Ahn. TweetMotif: Exploratory search and topic summarization for Twitter. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2010c.

Brendan O'Connor, David Bamman, and Noah A. Smith. Computational text analysis for social science: Model assumptions and complexity. In *Second Workshop on Computational Social Science and the Wisdom of Crowds (NIPS 2011)*, 2011.

Brendan O'Connor, Brandon Stewart, and Noah A. Smith. Learning to extract international relations from political context. In *Proceedings of ACL*, 2013.

Office of Management and Budget (USA). 2010 standards for delineating metropolitan and micropolitan statistical areas. *Federal Register*, 75(123), June 2010.

I. Ounis, C. Macdonald, and I. Soboroff. On the TREC blog track. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2008.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL*, 2013.

Bo Pang and Lillian Lee. *Opinion Mining and Sentiment Analysis*. Now Publishers Inc, July 2008. ISBN 1601981503.

John C. Paolillo. *Analyzing Linguistic Variation: Statistical Models and Methods*. CSLI Publications, 2002.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. English Gigaword fourth edition. *Linguistic Data Consortium*, LDC2009T13, 2009.

Michael Paul and Mark Dredze. Factorial LDA: Sparse multi-dimensional text models. In *Advances in Neural Information Processing Systems*, pages 2582–2590, 2012.

Michael Paul and Roxana Girju. A two-dimensional topic-aspect model for discovering multi-faceted topics. In *Proceedings of AAAI*, 2010.

Michael J. Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. In *Proceedings of ICWSM*, 2011.

Will D. Penny. Variational Bayes for  $d$ -dimensional Gaussian mixture models. Technical report, Wellcome Department of Cognitive Neurology, University College London, 2001. URL <http://www.fil.ion.ucl.ac.uk/~wpenny/publications/vbgmm.ps>.

Giovanni Petris. An R package for dynamic linear models. *Journal of Statistical Software*, 36(12): 1–16, 2010. <http://www.jstatsoft.org/v36/i12/paper>.

J. Piskorski and M. Atkinson. Frontex real-time news event extraction framework. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 749–752. ACM, 2011.

- J. Piskorski, H. Taney, M. Atkinson, E. van der Goot, and V. Zavarella. Online news event extraction for global crisis surveillance. *Transactions on computational collective intelligence V*, pages 182–212, 2011.
- Nicholas G. Polson, James G. Scott, and Jesse Windle. Bayesian inference for logistic models using Polya-Gamma latent variables. *arXiv preprint arXiv:1205.0310*, 2012.
- Keith T. Poole and Howard Rosenthal. *Congress: A political-economic history of roll call voting*. Oxford University Press, 1997.
- Hoifung Poon and Pedro Domingos. Unsupervised semantic parsing. In *Proceedings of EMNLP*, pages 1–10. Association for Computational Linguistics, 2009.
- Foster Provost and Tom Fawcett. *Data Science for Business*. O'Reilly Media, 2013.
- Kevin M. Quinn, Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209228, 2010.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL <http://www.R-project.org/>. ISBN 3-900051-07-0.
- Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. The ACL Anthology network corpus. In *Proc. of ACL Workshop on Natural Language Processing and Information Retrieval for Digital Libraries*, 2009.
- Anand Rajaraman and Jeffrey David Ullman. Mining of massive datasets. Cambridge University Press; <http://infolab.stanford.edu/~ullman/mmds.html>, 2011.
- Daniel Ramage, Christopher D. Manning, and Susan Dumais. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, page 457465, 2011.
- Lance Ramshaw, Elizabeth Boschee, Marjorie Freedman, Jessica MacBride, Ralph Weischedel, , and Alex Zamanian. SERIF language processing effective trainable language understanding. *Handbook of Natural Language Processing and Machine Translation*, pages 636–644, 2011.
- William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.
- Florencia Reali and Thomas L. Griffiths. Words as alleles: connecting language evolution with bayesian learners to models of genetic drift. *Proceedings. Biological sciences / The Royal Society*, 277(1680):429–436, February 2010. ISSN 1471-2954. doi: 10.1098/rspb.2009.1513. URL <http://dx.doi.org/10.1098/rspb.2009.1513>.
- Michaela Regneri, Alexander Koller, and Manfred Pinkal. Learning script knowledge with web experiments. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 979–988, 2010. URL <http://dl.acm.org/citation.cfm?id=1858681.1858781>.
- John R. Rickford. Ethnicity as a sociolinguistic boundary. *American Speech*, pages 99–125, 1985.

John R. Rickford. *African American Vernacular English*. Blackwell, 1999.

Margaret E. Roberts, Brandon M. Stewart, and Edoardo M. Airoldi. Structural topic models. 2013. URL <http://scholar.harvard.edu/bstewart/publications/structural-topic-models>. Working paper.

Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 2014.

Geoffrey Rockwell, Stéfan G. Sinclair, Stan Ruecker, and Peter Organisciak. Ubiquitous text analysis. *paj: The Journal of the Initiative for Digital Humanities, Media, and Culture*, 2(1), 2010.

Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1500–1510. Association for Computational Linguistics, 2012.

M. Rooth, S. Riezler, D. Prescher, G. Carroll, and F. Beil. Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, page 104111, 1999.

R.J. Rummel. The Dimensionality of Nations project. 1968.

Gerard Rushton, editor. *Geocoding health data: the use of geographic codes in cancer prevention and control, research, and practice*. CRC Press, 2008. ISBN 9780849384196.

Andrey Rzhetsky, Ivan Iossifov, Tomohiro Koike, Michael Krauthammer, Pauline Kra, Mitzi Morris, Hong Yu, Pablo Ariel Duboué, Wubin Weng, W. John Wilbur, Vasileios Hatzivassiloglou, and Carol Friedman. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics*, 37(1):43–53, 2004.

Adam Sadilek, Henry Kautz, and Jeffrey P. Bigham. Finding your friends and following them to where you are. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 723–732, 2012.

Evan Sandhaus. The New York Times Annotated Corpus. *Linguistic Data Consortium*, LDC2008T19, 2008.

A. Sanfilippo, L. Franklin, S. Tratz, G. Danielson, N. Miles, R. Riensche, and L. McGrath. Automating frame analysis. *Social computing, behavioral modeling, and prediction*, pages 239–248, 2008.

David Sankoff, Sali A. Tagliamonte, and Eric Smith. Goldvarb X: A variable rule application for Macintosh and Windows. Technical report, Department of Linguistics, University of Toronto, 2005.

Philip Schrodt and Kalev Leetaru. GDELT: Global data on events, location and tone, 1979-2012. In *International Studies Association Conference*, 2013.

Philip A. Schrodt. Automated coding of international event data using sparse parsing techniques. *International Studies Association Conference*, 2001.

Philip A. Schrodt. Twenty Years of the Kansas Event Data System Project. *Political Methodologist*, 2006.

Philip A. Schrodt. Precedents, progress, and prospects in political event data. *International Interactions*, 38(4):546–569, 2012.

Philip A. Schrodt and Deborah J. Gerner. Validity assessment of a machine-coded event data set for the Middle East, 1982-1992. *American Journal of Political Science*, 1994.

Philip A. Schrodt and Deborah J. Gerner. An event data analysis of third-party mediation in the middle east and balkans. *Journal of Conflict Resolution*, 48(3):310–330, 2004.

Philip A. Schrodt, Shannon G. Davis, and Judith L. Weddle. KEDS – a program for the machine coding of event data. *Social Science Computer Review*, 12(4):561 –587, December 1994. doi: 10.1177/089443939401200408. URL <http://ssc.sagepub.com/content/12/4/561.abstract>.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E.P. Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, 8(9):e73791, 2013. doi: 10.1371/journal.pone.0073791.

Ryan Shaw. Text-mining as a research tool. Duke Libraries, Text > Data series, September 2012. URL <http://aeshin.org/textmining>.

Stephen M. Shellman. Time series intervals and statistical inference: The effects of temporal aggregation on event data analysis. *Political Analysis*, 12(1):97–104, 2004.

Stephen M. Shellman. Coding disaggregated intrastate conflict: machine processing the behavior of substate actors over time and space. *Political Analysis*, 16(4):464, 2008.

Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of IEEE Symposium on Visual Languages*, pages 336–343. IEEE, 1996.

Aditi Shrikumar. *Designing an Exploratory Text Analysis Tool for Humanities and Social Sciences Research*. PhD thesis, University of California at Berkeley, 2013.

R. W. Sinnott. Virtues of the Haversine. *Sky and Telescope*, 68(2), 1984.

Aaron Smith and Lee Rainie. Who tweets? Technical report, Pew Research Center, December 2010. URL <http://pewresearch.org/pubs/1821/twitter-users-profile-exclusive-examination?src=prc-latest&#38;proj=peoplepress>.

Lauren Squires. Enregistering internet language. *Language in Society*, 39:457–492, September 2010.

Stan Development Team. Stan: A C++ library for probability and sampling, version 2.4, 2014. URL <http://mc-stan.org/>.

Seth I. Stephens-Davidowitz. The effects of racial animus on a black presidential candidate: Using Google search data to find what surveys miss. SSRN, 2012. URL <http://dx.doi.org/10.2139/ssrn.2050673>.

Benedikt Szmrecsanyi. Corpus-based dialectometry: a methodological sketch. *Corpora*, 6(1):45–76, 2011.

Matt Taddy. Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503):755–770, 2013.

Sali A. Tagliamonte. *Analysing Sociolinguistic Variation*. Cambridge University Press, 2006.

Sali A. Tagliamonte and Derek Denis. Linguistic ruin? LOL! Instant messaging and teen language. *American Speech*, 83, 2008.

Yla R. Tausczik and James W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 2009. URL <http://jls.sagepub.com/cgi/rapidpdf/0261927X09351676v1>.

Yee Whye Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 985–992. Association for Computational Linguistics, 2006.

Paul C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):11391168, 2007.

Mike Thelwall. Homophily in MySpace. *J. Am. Soc. Inf. Sci.*, 60(2):219–231, 2009. doi: 10.1002/asi.20978. URL <http://dx.doi.org/10.1002/asi.20978>.

Crispin Thurlow. From statistical panic to moral panic: The metadiscursive construction and popular exaggeration of new media language in the print media. *Journal of Computer-Mediated Communication*, 11(3):667–701, 2006.

Ivan Titov and Alexandre Klementiev. A Bayesian model for unsupervised semantic parsing. In *Proceedings of ACL*, 2011.

Ivan Titov and Alexandre Klementiev. A Bayesian approach to unsupervised semantic role induction. *Proceedings of EACL*, 2012.

Peter E. Trapa and Martin A. Nowak. Nash equilibria for an evolutionary language game. *Journal of mathematical biology*, 41(2):172–188, August 2000. ISSN 0303-6812. URL <http://view.ncbi.nlm.nih.gov/pubmed/11039696>.

Peter Trudgill. Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society*, 3(2):215–246, 1974.

John W. Tukey. *Exploratory data analysis*. Pearson, 1977.

Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *International AAAI Conference on Weblogs and Social Media, Washington, DC*, 2010.

Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. Where there is a sea there are pirates: Response to Jungherr, Jürgens, and Schoen. *Social Science Computer Review*, 30:235–239, 2012.

Peter D. Turney. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML)*, 2001.

Peter D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, page 417424, 2002.

Peter D. Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141188, 2010. ISSN 1076-9757.

Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald. The viability of web-derived polarity lexicons. In *NAACL-HLT: Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010.

Martin J. Wainwright and Michael I. Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers, 2008.

Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. *Advances in Neural Information Processing Systems*, 22:1973–1981, 2009.

Hanna M. Wallach. *Structured topic models for language*. PhD thesis, University of Cambridge, 2008.

Chong Wang and David M. Blei. Variational inference in nonconjugate models. *arXiv preprint arXiv:1209.4360*, 2012.

Wei Wang, David Rothschild, Sharad Goel, and Andrew Gelman. Forecasting elections with non-representative polls. *International Journal of Forecasting*, Forthcoming, 2014.

Martin Wattenberg and Fernanda B. Viégas. The word tree, an interactive visual concordance. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1221–1228, 2008.

Greg C. G. Wei and Martin A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704, 1990. ISSN 01621459. doi: 10.2307/2290005. URL <http://dx.doi.org/10.2307/2290005>.

William Wu-Shyong Wei. *Time series analysis*. Addison-Wesley, 1994.

Hadley Wickham. A layered grammar of graphics. *Journal of Computational and Graphical Statistics*, 19(1):328, 2010. doi: 10.1198/jcgs.2009.07098.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210, 2005.

J. Wilcox. Forecasting components of consumption with components of consumer sentiment. *Business Economics*, 42(4):2232, 2007.

Leland Wilkinson. *The grammar of graphics*. Springer, 2006.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language (HLT/EMNLP)*.

Benjamin P. Wing and Jason Baldridge. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 955–964. Association for Computational Linguistics, 2011.

Walt Wolfram and Natalie Schilling-Estes. *American English: Dialects and Variation*. Wiley-Blackwell, 2nd edition edition, September 2005.

Shaomei Wu, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Who says what to whom on twitter. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 705–714, 2011.

Eric P. Xing. On topic evolution. Technical Report 05-115, Center for Automated Learning and Discovery, Carnegie Mellon University, 2005.

Limin Yao, David Mimno, and Andrew McCallum. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 937–946. ACM, 2009.

Limin Yao, Aria Haghghi, Sebastian Riedel, and Andrew McCallum. Structured relation discovery using generative models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1456–1466. Association for Computational Linguistics, 2011.

Dani Yogatama, Michael Heilman, Brendan O’Connor, Chris Dyer, Bryan R. Routledge, and Noah A. Smith. Predicting a scientific community’s response to an article. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011.

Scott L. Zeger and M. Rezaul Karim. Generalized linear models with random effects; a gibbs sampling approach. *Journal of the American statistical association*, 86(413):79–86, 1991.

Menghan Zhang and Tao Gong. Principles of parametric estimation in modeling language competition. *Proceedings of the National Academy of Sciences*, 110(24):9698–9703, 2013.

Kathryn Zickuhr and Aaron Smith. 4% of online Americans use location-based services. Technical report, Pew Research Center, November 2010.

George K. Zipf. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Martino Fine Books, June 1949/2012.