Thesis Defense

New methods for large-scale analyses of social identities and stereotypes

Kenneth Joseph

CMU-ISR-16-107

June 8th, 2016

Institute for Software Research
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Dr. Kathleen M. Carley
Dr. Jason Hong
Dr. Lynn Smith-Lovin
Dr. Eric Xing

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

Support for the work presented in this thesis was provided, in part, by the Office of Naval Research (ONR) through a MURI N00014081186 on adversarial reasoning and the ONR through a MINERVA N000141310835 on State Stability. Work was also supplied by a Carnegie Mellon GuSH grant.

Keywords: Computational Social Science, Affect Control Theory, Natural Language Processing, Bayesian Networks

To anyone making it this far

Abstract

Social identities, the labels we use to describe ourselves and others, carry with them stereotypes that have significant impacts on our social lives. Our stereotypes, sometimes without us knowing, guide our decisions on whom to talk to and whom to stay away from, whom to befriend and whom to bully, whom to treat with reverence and whom to view with disgust.

Despite the omnipotent impact of identities and stereotypes on our lives, existing methods used to understand them are lacking. In this thesis, I first develop three novel computational tools that further our ability to test and utilize existing social theory on identity and stereotypes. These tools include a method to extract identities from Twitter data, a method to infer affective stereotypes from newspaper data and a method to infer both affective and semantic stereotypes from Twitter data. Case studies using these methods provide insights into Twitter data relevant to the Eric Garner and Michael Brown tragedies and both Twitter and newspaper data from the "Arab Spring".

Results from these case studies motivate the need for not only new methods for existing theory, but new social theory as well. To this end, I develop a new socio-theoretic model of *identity labeling* - how we choose which label to apply to others in a particular situation. The model combines data, methods and theory from the social sciences and machine learning, providing an important example of the surprisingly rich interconnections between these fields.

Acknowledgments

I am extremely grateful to my advisor, Kathleen Carley, for giving me a chance when (although she didn't know it) no one else would. She has also given me a deep appreciation of the connection between computational methods, the social and the cognitive sciences, and has allowed me room to grow as a scholar at this intersection, which I very much appreciate. Credit also goes to the members of my committee who were each helpful in shaping portions of this thesis with their recommendations and criticisms. I also acknowledge the social and scholastic support of several informal mentors. Michael Martin, Jürgen Pfeffer, Chul Gwon, Moises Sudit and Carol Frieze were instrumental in my development as a scholar, student and teacher, and I appreciate their mentorship more than they probably realize.

I have also had the privilege of making several lifelong friends and future colleagues at CMU that I could not have completed the work here or retained my sanity without - thank you especially to Geoff Morgan (and through Geoff, Jon Morgan), Wei Wei, Matt Benigni, Will Frankenstein, Jana Diesner, Hemank Lambda, Blase Ur and Manya Sleeper, but also to everyone else in CASOS and SC over the years. I also want to thank my first "colleagues", Anthony Ross and Murtuza Boxwala, for their continued support. More generally, I would like to thank all of my friends, from Buffalo to Bambi to the 'burgh, for supporting me even though they really don't know what the hell I do.

Finally, I would like to thank my mother, father and sister for not questioning my choices and for their unwavering commitment to my happiness and success. And, above all, I thank my wife, who has had to deal with me being mentally and/or physically absent from our life together for large stretches of the last six years. As this document makes quite clear, I have words for many things, but none for how much your love and support mean to me.

Contents

List of Figures

List of Tables

Chapter 1: Introduction

...thug today is a nominally polite way of using the N-word...It is a sly way of saying there go those black people ruining things again.
John McWhorter, on NPR, 4/30/2015

A social identity is a word or phrase used to define a particular type, group or class of individuals (Smith-Lovin, 2007). The identities we choose for ourselves or that are chosen for us impact our lives in important ways. For example, being labeled a woman or an African American can have a significant negative effect on your employment opportunities (Bertrand and Mullainathan, 2003). Identifying as LGBTQ can have negative effects as well- in 2010, *eight out of ten* American students who identified themselves as members of the LGBTQ community reported being bullied because of their sexual orientation[1].

As important as identities are to our social lives, however, fairly little is known about the universe of identities that can possibly be applied to a particular individual (Heise and MacKinnon, 2010). That is, while we understand that identities are important, what words actually make up this set of "identities" is relatively unknown. Chapter 3 of this thesis describes **a new method to extract social identities used by a population, with a particular application to Twitter data** (Joseph et al., 2016b). The model we develop significantly outperforms baseline approaches on this task by leveraging modern approaches to natural language processing on Twitter. After developing and validating our model to perform this task, we embark on a case study of identity usage in a population of 250,000 geotagged Twitter users who were central to the discussion of the Eric Garner and Michael Brown tragedies, as well as a set of 156,000 Twitter users actively engaged in discussions of the Arab Spring. These case studies reveal interesting properties about the universe of identities available to these two populations, how these identities cluster semantically and how users feel about identities within these clusters.

Understanding the universe of identities available to a particular population is useful on its own; we are able to see what labels could possibly be applied to particular individuals and how labels are related. However, we are still left after Chapter 3 with the question of *why* identities so strongly impact our lives and *how* they are appropriated to particular individuals in particular situations. The answers to these two questions lie in the *stereotypes*, or meanings, attached to each identity. The scientific study of stereotypes dates back at least 80 years, with Katz and Braly's (1933) work exploring racial stereotypes of college-aged students representing some of the earliest research I am aware of in the area. This thesis focuses on how stereotypes are *quantified*, that is, how they are defined mathematically and how they are measured based on these definitions. A significant amount of literature spanning social psychology (Burke, 1980; Fiske et al., 2002;

[1] http://www.pacer.org/bullying/about/media-kit/stats.asp

Heise, 2007; Rogers et al., 2013), cognitive psychology (Freeman and Ambady, 2011; Kunda and Thagard, 1996; Read and Miller, 1998; Schröder and Thagard, 2014), and neuropsychology (Cikara and Van Bavel, 2014; Van Bavel and Cunningham, 2010), among other fields, focuses on the quantification of stereotypes.

From this work, two general problems plaguing the measurement of stereotypes can be surmised. First, even with a particular definition of stereotypes in mind, stereotypes are difficult and expensive to measure. Nearly all existing methods of measuring stereotypes rely on survey data, which are notoriously difficult to collect. Chapter 2 develops a **new method to extract stereotypes from newspaper data** (Joseph et al., 2016a). From a social science perspective, the method I develop can provide estimates of stereotypes in days, while surveys may take months or even years to complete. My approach also can, unlike surveys, be applied to historical data. From a computational perspective, I develop a new means of concept-level sentiment analysis using subject-verb-object triplets that is complementary to existing approaches. I develop an efficient Gibbs-EM algorithm to infer model parameters and show that the model is competitive in a prediction task against various baseline models. After developing and validating the approach, I use it for a case study on a corpora of 700,000 English newspaper articles related to the Arab Spring. Amongst other observations, I find that the Muslim identity was largely cast as the victim rather than the villain in my data. In contrast, I observe that Sunnis, the majority sect of Islam in the Arab world, may have been villianized by the media, a finding that can be explained in various ways but that nonetheless matches what was occurring on the ground at the time.

The second, and perhaps more vexing, problem plaguing the measurement of stereotypes is that existent measurement models are difficult to contrast and compare, and hence difficult to leverage in tandem to glean a better understanding of stereotypes. While w models within disciplines show fairly similar properties (see Cikara and Van Bavel, 2014; Freeman and Ambady, 2011; Rogers et al., 2013, for reviews of social psychological, cognitive psychological and neuropsychological models, respectively), across disciplines models vary on their faithfulness to different principles. Cognitive models assume sociological processes away to researcher intuition or chance, while social psychological models provide strong sociological foundations and skirt properties of cognition. Fixed, intuition-based parameters of one model are therefore the same parameters of another that are estimated from data. Even where parameters of interest are similar, data sources used to construct models varies tremendously, from "big data" and text analytics (Bamman et al., 2014; Heise and MacKinnon, 2010) to a combination of survey data and literature review (Freeman and Ambady, 2011; Schröder and Thagard, 2014).

Chapter 4 **develops a new approach to measure stereotypes in a population of Twitter users**. From a socio-cognitive theoretic standpoint, the model I develop of stereotypes is an *attributed network model* that combines an existing cognitive model of stereotypes with an existing social psychological model of stereotypes. The cognitive portion of the model draws on *parallel constraint satisfaction models* (Freeman and Ambady, 2011; Kunda and Thagard, 1996), which focus largely on *semantic relationships* between identities (e.g. the role relationship between "brother" and "sister"). The social psychological portion of the model draws on *Affect Control Theory* (Heise, 1987, 2007; Robinson et al., 2006), which focuses on how people *feel* about particular identities.

My work is the first attempt I am aware of to combine semantic model of stereotype from the cognitive sciences with an affective model of stereotypes from the social psychological literature.

The approach I develop to infer parameters for the semantic portion of the model is based on recent literature (Chen et al., 2012) for efficient Gibbs-sampling based inference of the Correlated Topic Model (Blei and Lafferty, 2007), from which our model is derived. The approach I develop to infer parameters for the affective portion of the model is a generalization of the approach described in Chapter 2 that is able to incorporate a wider scope of sentiment information from tweets into the inference process. In sum, the inference algorithm is an efficient, parallelized Gibbs sampling model fully implemented on Apache Spark.

After showing that the model presented outperforms competitive baselines in a prediction task on held-out data, we embark on a case study of a set of 45K Twitter users interested in the Eric Garner and Michael Brown tragedies, a subset of the users in the dataset from Chapter 3. I find, among other things, that users in our dataset had more negative feelings toward police and more positive feelings towards protesters than prior information we have on these identities would have led us to believe. This finding matches well with aspects of the coverage of these tragedies. I also show that differentiating between affective and semantic stereotypes allows a more nuanced view of stereotypes of the identity thug than existing "deep learning" approaches to inferring word meaning. More specifically, I find that it may be easy for (what we consider to be) undesirable stereotypes to "creep in" to existing computational models of word meaning, and that modeling the affective meaning of words may be an important step in mitigating this process. Finally, I also find more concrete evidence that identities are clustered into hierarchically organized "institutions", that is, that semantic stereotypes of identities reveal hierarchically structured groups of identities that are related by the settings in which they are generally observed.

Like much of the research in social psychology, Chapters 2 and 4 focus on how a person with a particular label will be stereotyped. In Chapter 5, I use what I have learned from this work to ask what is essentially the opposite question - how do the stereotypes attributed to a particular identity inform how it will be applied to particular individuals in particular social situation? I refer to this question as the *identity labeling problem*, and in Chapter 5 **I develop a new mathematical model of the identity labeling problem**.

I develop this model in three stages. I first construct a generic representation of the identity labeling problem that can be used to represent several other similar models, including Hoey et al.'s (2013a) BayesACT. I then develop a simple, parsimonious model of identity labeling that focus on three different kinds of relationships between identities that have been shown to be important in the identity labeling process in prior work. The first type of relationship is *semantic similarity*, which defines the extent to which two identities can be used to define the same person. The second is *semantic association*; how likely it is that two identities are seen in the same social context. Finally, fitting with the previous chapters, I consider *affective similarity*.

I run two surveys to test how each of these factors impact that way that survey respondents labeled individuals in hypothetical situations, finding that semantic factors have universally strong impacts on how survey respondents label people, but that affective similarity may only matter in certain situations. I discuss possible reasons for this and how it relates to prior work on identity labeling from Heise and MacKinnon (2010). These surveys also suggest that most identity relationships can be characterized into one of four types - either the identities are entirely unrelated (e.g. "dentist" and "colonel" , are *counter-identities* (Burke, 1980) that are commonly found within the same social context but refer to different individuals (e.g. "brother" and "sister"), are culturally synonymous (e.g. "doctor" and "physician") or are *multiple identity pairs*, meaning

they are rarely used in the same social context but are, counterintuitively, often used to refer to the same individual (e.g. "man" and "brother"). I consider how this typology informs our understanding of identity relationships.

Finally in Chapter 5, I construct one last identity labeling model that goes beyond these three high-level concepts to considering a characterization of *how* identities are semantically associated, semantically similar and/or affectively similar. I give initial results that suggest the model I develop is useful for inferring things like implicit racialization of particular identities (e.g. that survey respondents often implicitly associated the identities "criminal" and "black person") and the relationship between different axes of affect and semantic meanings, e.g. the relationship between the implicit gender of an identity and its goodness/badness on an affective scale.

Each part of this thesis makes a major contribution to the quantitative study of identity and stereotypes. Chapter 3 introduces a new method to infer the universe of identities available to a particular population and how these identities cluster semantically. Chapter 2 develops a new way to extract affective stereotypes of a particular set of identities from newspaper data. Chapter 4 develops a new approach to jointly infer affective and semantic stereotypes from Twitter data. Chapter 5 provides a new quantitative socio-theoretic model of how identities are applied to individuals in social situations and infers parameters of it from survey data.

This thesis also makes three important contributions to the field of natural language processing. Chapter 2 provides a new approach to concept-level sentiment mining that, as Chapter 4 shows, may be usefully incorporated into existing models. Second, Chapter 4 also gives further evidence (e.g. Maas et al., 2011) that future NLP methods may do well to incorporate affective information into measurements of semantic similarity. Finally, along these same lines, Chapter 5 provides sociological explanations for what I perceive to be various flaws in existing data used to evaluate the ability of NLP models to measure association and similarity. I provide a new dataset of surveys that may be useful in future evaluations.

In addition to these contributions, this thesis presents proof that computational social science does not have to be done by applying an existing computational model to a problem and interpreting its results to push forward social science. Rather, in this thesis the I develop new methods that are explicitly grounded in socio-cognitive theory. This approach allows me to take the results of these methods and understand not only where the methods I develop can be improved, as a traditional machine learning researcher would, but also where the socio-cognitive theory of identity and stereotype is deficient. As will become clear over the course of this document, Chapters 2, 4 and 5 represent an iteration of this process by which contemporary social science drives the problems I tackle and the models I develop, and contemporary approaches to machine learning using graphical model formalisms help me to refine that theory. I believe this approach to computational social science is relatively underserved in the literature but is extremely useful in identifying existing questions in each field.

Chapter 2: A Model to Extract Affective Stereotypes from Newspaper data

2.1 Introduction

Let us define a *social event* in the sense of Heise (2007) as a situation in which an *actor* enacts a *behavior* on an *object*. Further, let us assume that both the actor and the object are *identities*, which we will define as nouns that are commonly used to allude to a social category (Tajfel and Turner, 1979). Finally, we assume that each identity has a particular affective meaning, or sentiment.

In theory, an infinite number of social events could occur between two identities on an everyday basis. For instance, there are few concrete barriers that prevent all strangers that pass each other on the street from shaking hands. In practice, however, there are many constraints on the social events we are willing to engage in and those we will observe in our everyday lives. Some of these constraints are hard, or physical. Geospatial distance, for example, acts as a barrier that restricts the types of identities that come in contact. Others are "soft", existing within our perceptions of cultural norms. These soft constraints are particularly interesting, as they passively define the "right" way to interact without any actual physical restrictions.

For example, assume that you are a new elementary-school teacher. It is unlikely that your first action will be to "beat up" your students, even though you are perfectly capable of doing so. Rather, you would be considerably more likely to, for example, "advise" them. "Advising" is an act that fits your, and almost everyone's, intuitions for the identity of a teacher, the identity of a student and the relationship between them. Now consider the situation where you observe a policewoman roughly handling a suspect. Depending on your views on the police, your cultural upbringing and the context in which you observe this act (among other things), your perception of this event may range from an actor carrying out one's duty as an officer to purely inhumane behavior. Thus while some soft constraints, such as those imposing our views on teachers and children, are almost universal (at least within a particular culture), the variability in our emotional response to other events suggests the incredible complexities that can arise in understanding how an individual perceives and engages in social events.

The present work is interested in developing a methodology that allows for a better understanding of how one particular form of these soft constraints, *affective* constraints, mediate perceptions of a particular set of identities engaging in a particular set of behaviors across many social events. Specifically, we develop an approach that is able to infer affective meanings, or sentiments, of identities and behaviors from a large text corpus. These affective meanings, we will show, serve as strong constraints on perceptions of social events within the corpus. In pur-

suing such a method, three chief issues must be overcome.

Issue 1. While there are an increasing number of databases and tools for extracting world events (e.g. GDELT; Leetaru and Schrodt, 2013) and social behaviors of individuals (e.g. social media, mobile phone records), there is a surprisingly limited amount of data and computational methodologies supporting the extraction of social events engaged in by the generalizable social identities of interest (e.g. "teacher"). In the present work, we present a partial solution to this problem. We first use dependency parsing (Kübler et al., 2009) of newspaper data to extract social events. We then manually clean the resulting output to pull out interesting identities and behaviors from the noisy result of the dependency parse. While we are far from the first to use dependency parsing to extract events from text (for a recent example, see O'Connor et al., 2013), few, if any, have considered the goal of extracting events with the aim of using them to infer affective characteristics of identities.

Issue 2. In order to make use of these extracted events, we must then address the issue of how to model the affective constraints that restrict actions in and perception of social events. In the present work, we use Affect Control Theory (ACT) (Heise, 1987, 2007; Robinson et al., 2006), which provides a formal social psychological model that describes the following (among many other things):

- The dimensions of sentiment along which we perceive identities and behaviors
- How social events change our perceptions of others
- How we engage in and think about social events in a way that confirms our sentiments

ACT is a "control" theory in that it assumes humans seek to *maintain* preexisting, culturally shared perceptions of identities and behaviors in transient impressions that are generated when social events are observed or carried out. While we may try to maintain these meanings through various methods, our efforts are all carried out in an attempt to reduce the *deflection*, or difference, between our impressions of individuals and culturally-shared sentiments of the identities they represent. ACT assumes that events we expect, or that we are more willing to carry out, are generally low in deflection, as these events are easy to incorporate into our current world-views. For example, the statement "the teacher advises the student" has been estimated to have a deflection of approximately 0.8, while the statement "the teacher beats up the student" has a deflection of 15.4[1].

In ACT, the deflection of a social event is estimated based on two sources of data. First, ACT scholars maintain a large database of survey results that serve as estimates for culture-wide sentiments of a host of identities and behaviors[2]. These sentiments are defined within a three-dimensional affective latent space that has been both theoretically and empirically validated (Osgood, 1975). Second, ACT scholars have developed a set of *change equations* which mathematically describe how the observation of a particular social event changes our perception of the actor, behavior and object involved in it. Given the position of these three entities in the affective space and a change equation, the (unnormalized) Euclidean distance between the affective meanings of the entities before versus after the event defines the level of deflection for the event. We can thus use deflection to understand the relative likelihood of different social events, implic-

[1] these values were computing using the INTERACT Java program (Heise, 2010a) with the Indiana 2002-2004 dictionary (Francis and Heise, 2006)

[2] `http://www.indiana.edu/~socpsy/ACT/data.html`

itly giving us an understanding of the affective constraints imposed within the social system of interest.

Unfortunately, while ACT's dictionaries already encompass thousands of identities and behaviors, the data within them are difficult to apply directly to a specifically themed corpus. While collecting new data on new identities and behaviors of interest is possible, it currently requires lengthy survey procedures. Additionally, ACT makes the tenuous assumption that point estimates are sufficient to describe the affective meanings of identities and behaviors (Hoey et al., 2013a). Finally, while the theory has been tested using survey methodology with individuals in several large cultural groups (e.g. nations), how best to identify any possible differences within these cultures without additional surveys remains an open question. ACT thus holds the potential to be used to provide insight into the affective constraints that shape social events and their perceptions. However, both methodological and data issues prevent a direct application of the theory in many settings of interest to scholars.

Issue 3. The third issue at hand is thus bow best to adapt the concepts involved in ACT into a model that can overcome, at least in part, it's current limitations. The primary contribution of the present work is a probabilistic graphical model (Koller and Friedman, 2009) that provides an initial and substantial step forward in this direction. The model we introduce has four desirable features in that it:

- infers affective meanings for identities and behaviors not currently in the ACT dictionaries
- incorporates prior knowledge from existing ACT dictionaries
- infers where multiple "senses" of a particular identity or behavior exist within our dataset
- provides a variance for the sentiment of each sense of each identity and behavior

Our approach is the first effort we are aware of to apply ACT concepts in an automated way to full text data. The statistical model we develop can be applied in a semi-automated fashion to any text corpus from which social events can be extracted, making it a potentially useful tool across a host of sociological domains. Further, from a natural language processing (NLP) perspective, while many other approaches exist to extract sentiment from text (see, e.g. Pang and Lee, 2008), such approaches typically exist on a single "good/bad" dimension. In comparison, our use of ACT allows for a multi-dimensional approach to understanding sentiment in text. This approach is critical in fully interpreting perceptions of identities, behaviors and social events (Osgood, 1975).

After describing the inner workings of our model, we provide a case study on a set of social events extracted from a corpus of approximately 600,000 newspaper articles relevant to the Arab Spring. After rigorous cleaning, the dataset contains 102 identities and 87 behaviors of interest that engage in 10,485 social events over the span of 30 months. Of the 189 identities and behaviors, only 84 (44%) exist in the original Affect Control dictionaries. Thus, we obtain new EPA profiles for many of the important identities and behaviors of interest in our dataset, and can use these to better understand how the English speaking news media perceived these identities and behaviors as the social movement evolved.

Naturally, our understanding is limited by the quality of model output. To this end, we provide a rigorous quantitative analysis of the effectiveness of the model on the task of predicting the behavior one identity enacts on another. Our model's performance improves over several baseline approaches on the prediction task, though struggles with issues of data sparsity in comparison to the strongest of our baselines. Still, the final model we present gives meaningful

affective meanings for the identities and behaviors within the dataset, which none of the baseline models are able to provide.

As the quantitative analysis suggests that the model fits the data reasonably well, we also (cautiously) consider the implications of model output on our understanding of news media coverage of the Arab Spring. Most interestingly, we observe a discrepancy in the way major English-speaking news outlets portrayed the generic Muslim identity as opposed to the more specific Sunni and Islamist identities.

2.2 Related Work

In this section, we first provide background on Affect Control Theory. As our methodology also draws comparisons to a variety of other tasks in the NLP literature, we also touch on efforts in this domain, in particular existing approaches to sentiment mining.

2.2.1 Affect Control Theory

Overview

Affect Control Theory, originally introduced by Heise (1979, 1987, 2007), presents a compelling model of how perceptions of the affective meaning of identities and behaviors develop. ACT also details how these perceptions can simultaneously exist for social categories and generic behaviors as well as for specific individuals and individual behavioral acts. In detailing these processes, ACT uses a host of ideas beyond the aforementioned concepts of identities, behaviors, the change equation and deflection. We discuss here only the portions of the theory relevant to our model, which center mainly around these four basic components. For those interested in a more complete discussion, we refer the reader to the chapter by Robinson et al. (2006) and the book by Heise (2007).

For matters of convenience, we will use the term *entities* where we are discussing something that applies to both identities and behaviors. All entities have an affective, or sentimental, meaning in a three dimensional latent space, the dimensions of which draw from early work by Osgood (1975) on measuring numeric profiles of affective meanings. The first dimension is evaluative, which describes the "goodness" or "badness" of an identity or behavior. The second dimension is potency, which describes the "powerfulness" (weakness) of an entity. The final dimension is activeness, which defines the level of energy or excitedness of a given entity. Each dimension is defined on the continuous interval $[-4.3, 4.3]$.

Combined, these three dimensions form what can be referred to as the EPA space. Within EPA space, all entities hold a particular position that defines their *EPA profile*. For example, the EPA profile of the identity "teacher" is (0.72, 1.87, 1.41), indicating that teachers are relatively good, powerful and active. In contrast, the EPA position of a student is (1.49, 0.31, 0.75), which shows students are "more good" than teachers, but much less powerful and active[3]. EPA positions of these entities, and many others, have been estimated by Affect Control researchers through a vast collection of survey experiments run across individuals from a host of cultures[4]. These values define the *fundamental*, culturally-shared meanings of identities and behaviors.

[3] Values from the Indiana 2002-2004 sentiment dictionary Francis and Heise (2006)

[4] The methodology involved in these surveys has evolved over time, and a thorough discussion can be found in (Heise, 2010b)

ACT assumes that the EPA profile for an individual instantiation of an entity may differ from the generic entity's EPA position. Thus, one may perceive a particular teacher as being "less good" than teachers in general. While in general fundamental meanings are assumed consistent throughout a culture, the theory also allows for the possibility that two different people may have different perceptions of the EPA profile of a generic identity or behavior. Variations of this sort are generally assumed to occur at the boundaries of social groups and social institutions. For example, Smith-Lovin and Douglas (1992) show that individuals in a gay, religious institution had uniquely positive views of the gay, cleric and congregation identities. As the authors state, these individuals "transformed both religious and gay identities so that the homosexual person [could] participate in religious rituals while not abandoning his or her gay identity".

Thomas and Heise (1995) provide a broader exploration of these multiple senses, showing that systematic differences do exist across social groups (e.g. gender) and via the extent to which individuals are embedded in multiple networks. The model in the present work only partially deals with the fact that different sentiments of the same entity may exist across social groups. On the one hand, the model we use allows for the possibility of multiple perceptions of the same generic identity. However, we also assume that one of the various possible perceptions for each entity in our event data is representative of an American cultural standpoint as provided by a particular ACT dictionary.

In addition to defining culturally shared sentimental meanings of entities, ACT also defines how one's perception of, for example, an individual teacher may develop as the teacher is observed carrying out different social events. The perception of a particular actor, behavior and object that we have before an event is known as the *pre-event transient*. The *post-event transient* describes our perception of the entities in the event after the event has been completed. In general, social events can be chained together such that transient impressions of a previous event become pre-event impressions for the next. In the present work, however, we will assume that each event occurs in isolation, and therefore that the pre-event impression are equal to their fundamental meaning.

The changes in impressions due to a particular social event are calculated in ACT using a change equation, which gives the post-event impression from a function of the pre-event one. The change equation mathematically defines the intuitive way in which pre-event impressions are altered by the social event that is observed. For example, a teacher should be seen as "less good" after beating up a child, and beating up should also be seen as less bad of an action. ACT postulates that the greater the difference, or *deflection*, between culturally-shared, fundamental impressions and post-event impressions is, the less likely an event is to occur. Thus, ACT postulates that we "prefer" to perceive and engage in social events in a way that aligns with our fundamental beliefs.

Affect Control researchers have used survey data to estimate the form of and parameters for the change equation (Heise, 2007; Smith-Lovin, 1987b). They have found that the form of the change equation may differ depending on national culture. In the present work, we assume for computational purposes that there exists only a single, universal change equation. This assumption, while still likely flawed, is supported by recent work that suggests differences in change equations across cultures may be due at least in part to weaknesses in earlier estimation techniques rather than to differences in the data (Heise, 2014).

Mathematical Model

Having given an overview of ACT, we now turn to the mathematical model given by the theory. To do so, we first introduce the form of the pre-event transient vector for a social event. Equation (2.1) gives this vector, which contains the EPA profiles associated with the three entities (*a*ctor, *b*ehavior, *o*bject) in a social event.

$$f = \begin{bmatrix} a_e & a_p & a_a & b_e & b_p & b_a & o_e & o_p & o_a \end{bmatrix} \tag{2.1}$$

Given the form of this vector, we can now describe how a social event changes these pre-event impressions to produce a post-event impression. This change occurs via the application of the change equation to a particular vector f. Though a variety of estimation methodologies have been used to estimate the change equation (Heise, 2007, 2014), the form of the equation is expected to define a polynomial, multiplicative combination of the pre-event transients. Thus, we can represent the change equation with two parts. First, the function $G(f)$ gives the subset of terms in the power set of f that have been estimated to impact the formation of the transient. As of the writing of this article, $G(f)$ is the following:

$$\begin{aligned} G(f) = [1 \quad & a_e \quad a_p \quad a_a \quad b_e \quad b_p \quad b_a \quad o_e \quad o_p \quad o_a \quad a_e b_e \quad a_e o_p \quad a_p b_p \quad a_a b_a \\ & b_e o_e \quad b_e o_p \quad b_p o_e \quad b_p o_p \quad a_e b_e o_e \quad a_e b_e o_p] \end{aligned} \tag{2.2}$$

Second, for each element of $G(f)$, we can define a set of coefficients, M, that describes the extent to which the element modifies the value of each element in f. The matrix M is thus a two-dimensional matrix with $|f|$ rows and $|G(f)|$ columns. The $M_{i,j}$ element of M describes the extent to which the jth coefficient of $G(f)$ impacts the ith element of the transient.

The deflection of a particular event is a measure of the squared Euclidean difference between the post-event transients and fundamental impressions. Because the pre-event impression is set equal to the fundamental, we can equivalently define deflection as the squared Euclidean difference between the pre and post-event transient impressions, as shown in Equation (2.3). In the equation M_{i*} is the ith row of M.

$$Deflection = \sum_{i}^{9} (f_i - M_{i*}^T G(f))^2 \tag{2.3}$$

It is important to note that because of the way the deflection equation is constructed, one can reassemble it as a quadratic function of the form $c_0 f_i^2 + c_1 f_i + c_2$ for any single element of f, f_i, if all other elements of f are considered to be constant. That is, if we were to actually replace M_{i*} and $G(f)$ with the regression model provided by ACT scholars, perform the multiplication of the squared term, all addition and all simplifications possible, we would end up with a long expression consisting of linear and quadratic combinations of all elements in f, plus a constant (e.g., $1.3 + .5 f_1 f_2^2 + .3 f_5^2 + ...$).

If we treat all $f_j, j \neq i$ as known (as constants), then this massive quadratic equation will reduce to the three term quadratic equation above with constants c_0, c_1 and c_2. The values of c_0, c_1 and c_2 can be computed using the equations above and will consist of nonlinear combinations of constants, including those elements of f, $f_{j,j \neq i}$, that we assume constant. This observation is

vital in developing the Gibbs sampling equations for our model[5].

Bayesian Affect Control Theory

Recently, Hoey and colleagues (Hoey et al., 2013a,b) converted aspects of ACT's mathematical model into one piece of a Partially Observable Markov decision process (POMDP). Their POMDP is used to train an intelligent tutoring tool, and thus their efforts are in a distinctly different vein. However, insights from the their efforts are directly relevant to our model. Most important, perhaps, is Hoey et. al's observation that one can exponentiate and negate the deflection equation to produce a true probability distribution. In doing so, a rearranging of the terms produces a multivariate normal distribution that makes Bayesian analysis feasible. While our model uses substantially different techniques, the relationship between the exponentiated form of the deflection equation and the normal distribution also plays an important role in the development of the model.

2.2.2 Other related approaches

The extraction of the sentimental meaning of different terms in a text is far from novel in the NLP community. Such efforts typically fall under the domain of sentiment analysis, defined as the extraction of emotional content from text, often in combination with other forms of data suitable for machine learning approaches. For a slightly dated but still very much relevant review of sentiment analysis techniques, we refer the reader to (Pang and Lee, 2008). In general, our approach differs in two important ways from previous sentiment mining approaches. First, there exists only a single other previous work that uses data made available by ACT researchers. Ahothali and Hoey (2015) apply an ACT-based model to social events extracted from news headlines. Their work differs in that they use only news headlines, use manual coding (via Mechanical Turk) to extract social event structure from text rather than the semi-automated approach defined here, and only extract a single EPA profile for each entity. Second, and perhaps most importantly, our approach moves beyond sentiment analysis tools that extract sentiment along single, evaluative dimension. Instead, we place identities and behaviors into a more empirically consistent three-dimensional latent, affective space. Beyond the work of Ahothali and Hoey (2015), few efforts have been made in this direction in the NLP community.

While unrelated to sentiment analysis, our use of the affective latent space draws comparisons to techniques like Latent Semantic Analysis (LSA) (Deerwester et al., 1990) and more recent approaches involving neural networks and "deep learning" (Lee et al., 2009) that place words into latent spaces that are representative of their meaning. Such approaches have been shown to be useful in both understanding meaning and in prediction problems. For example, recent convolutional neural network models have been developed that are able to solve analogies via simple algebra and distance models (Mikolov et al., 2013b), not unlike the methods for finding optimal behaviors for social events in ACT that we will describe below.

Finally, existing NLP tools, perhaps most notably Latent Dirichlet Allocation (LDA) (Blei et al., 2003), have formalized many of the difficulties involved with Bayesian analysis of text data and have shown the effectiveness of considering terms as belonging to many latent "topics"

[5]Note that while one could, quite simply, provide a regression model in which either the constant c_0 or c_1 is zero, in practice such an occurrence is unlikely and in any case it can be shown via simple algebra that the equations used here do not fit this case.

(or in our case, entities having multiple latent "senses"). The most relevant of these models is the work of O'Connor et al. (2013), who infer classes of behaviors that countries enact on each other over time. The authors use dependency parsing to extract events in which one country enacts a behavior on another. They then develop a model that jointly infers types of behaviors between countries and the extent to which the relationship between different countries is described by these classes of behaviors.

2.3 Extracting Social Events from Text

Country	Num. Articles	Country	Num. Articles
Algeria	11,059	Bahrain	21,314
Egypt	111,779	Iran	138,343
Iraq	101,147	Jordan	23,060
Kuwait	14,559	Lebanon	35,071
Libya	92,101	Morocco	27,153
Oman	8,581	Saudi Arabia	59,406
Syria	96,893	Tunisia	28,485
United Arab Emirates	73,029	Yemen	21,146

Table 2.1: Countries of interest to the present work and number of newspaper articles relevant to them

In order to use Affect Control Theory, we require a set of social events engaged in by entities of interest. In the present work, we are interested in understanding news media perceptions of the Arab Spring. In order to extract the requisite social events, we rely on a corpus of approximately 600K newspaper articles that we have collected. This large news corpus provides valuable information about numerous social events throughout the Arab Spring, and analyses of these texts can provide insight into behavior Joseph et al. (2014). The newspaper articles were extracted from LexisNexis Academic's corpus of "Major World Publications"[6] and were written between July 2010 and December 2012. We only extract articles written in English, and only consider articles that LexisNexis has indexed using its proprietary algorithms as being relevant to one or more of sixteen countries involved, either directly or tangentially, in the Arab Spring. These countries, and the number of articles relevant to them, are listed in Table 2.1[7].

Extraction of social events from text requires extracting information about "who did what to whom". While we expect such social events to be rampant in the text, the extraction of this type of information is an area of on-going research in the NLP community and has been studied within the subdomains of both dependency parsing (Kübler et al., 2009) and semantic role labeling (Carreras and Màrquez, 2005). Here, we use dependency parsing, as methodologies for dependency parsing are more readily available for the type of data we use. Specifically, we use the Stanford CoreNLP pipeline (Manning et al., 2014) to perform dependency parsing on our full set of data with the recently implemented, state-of the art model-based parser (Zhu et al.,

[6] `http://www.lexisnexis.com/hottopics/lnacademic/?verb=sf&sfi=AC00NBGenSrch&csi=237924`

[7] Note that a news article may be indexed by LexisNexis as being relevant to more than one country; hence the

2013). For more details on the general techniques and ideas behind dependency parsing, we refer the reader to (De Marneffe and Manning, 2008; Kübler et al., 2009). For an online example of dependency parsing, visit `http://nlp.stanford.edu:8080/parser/index.jsp`.

Quite simply, dependency parsing uses a variety of statistical techniques to extract from each sentence in our corpus the ways in which different terms are linguistically dependent on others. We run the dependency parser on all sentences from our corpus and extract all relations where we find both the subject and direct object of a verb. The subject, verb and object of the dependency parse are lemmatized[8] to their normalized form and then output for further processing.

This procedure allows us to extract social events from the text. For example, from the sentence "The teacher advised the student" the dependency parser (and post-parsing lemmatization) would extract the relationship "teacher advise student". Naturally, this process also extracts a host of noun-verb-noun relationships that are *not* social events, i.e. cases where either of the nouns are clearly not identities ("sanction help talk"), cases where the behavior is ambiguous ("husband say wife") and cases where the dependency parser appears to simply get confused ("issue hold talk"). To filter these events out from the data as best we can, we use a two-pass approach to cleaning. The first pass engages a variety of heuristics to remove highly irrelevant results. We then use a second, manual pass to further increase the relevancy of our data.

We use five heuristics in our first pass cleaning over the data. First, we ignore any events which do not contain at least one ACT identity or behavior. Though this is not required for our model, we find this serves to remove a host of uninteresting dependency parsing outputs. Second, we remove from the data any events appearing in highly similar sentences on the same day. This acts as a crude form of shingling (Rajaraman and Ullman, 2011), which helps in ensuring that we do not double count events from articles that contain nearly the same exact content reiterated by different outlets (O'Connor et al., 2013). Two sentences are similar if the noun-verb-noun relation extracted, along with any terms dependent on these three words (e.g. adjectives) are the exact same. Third, we ignore any relations where the subject or object is a pronoun. Such relations may be useable in the future if co-reference resolution is performed (Soon et al., 2001), but due to the computational complexity of doing so and the relatively high level of noise this process tends to induce, we do not use it at this point. Fourth, we ignore any results in which we observe the term "not" before the verb of the dependency parse, as we were unsure how to use ACT equations under this negation. Finally, we ignore any behaviors and identities that appear less than 25 times in our dataset, recursively removing events until all terms satisfy this requirement.

After this first pass of cleaning, we were left with approximately 5300 possible identities, 1300 possible behaviors and approximately 1.3M (of 1.7M total) social events. At this point, it was feasible to manually validate that all remaining entities extracted from the dependency parsing were indeed things we considered to be identities or behaviors, even if we could not consider the feasibility of each event individually. While future work may allow for more noise in the data, the present work was chiefly focused on model development, and thus we err on the

values in Table 1 sum to a value greater than 600,000

[8]Lemmatization is a process by which words are normalized in a deterministic fashion to facilitate analysis. Lemmatization includes stemming (e.g. changing "walking" to "walk" but also other steps, like synonym replacement, (e.g. replacing "better" with "good"). It is standard practice in the NLP literature to perform lemmatization before analysis

side of caution in deciding whether to include or exclude entities. For identities, we included only terms in the current ACT dictionaries, terms representing national identities and/or governments (including national leaders), well-known social groups and general identities that were deemed to be interesting by the majority of co-authors (e.g. "protestor"). We only included behaviors that were unambiguous in the action being taken and that could feasibly be expected to relate two identities. For example, we chose not to use the term "have" as a behavior, as it is unintuitive to consider one identity "having" another.

After finishing this processing, we again ensured that all identities and behaviors in our data occur at least 25 times in the cleaned dataset to ensure there was enough data to provide a reasonable estimate of their EPA profile. The final dataset we use has 102 identities, 87 behaviors and 10,485 social events.

2.4 Model Description

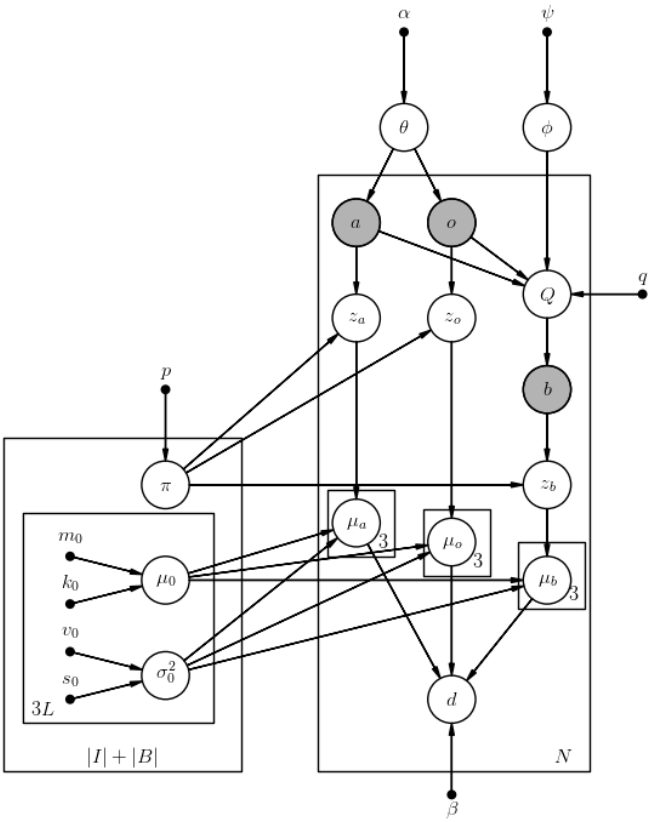


Figure 2.1: A depiction of the probabilistic graphical model used in the present work using standard plate notation

Figure 2.1 depicts the probabilistic graphical model used in the present work using standard

Variable	Description
n	Each social event, n, consists of the triple $< a_n, b_n, o_n >$
a_n, o_n, b_n	The actor, object and behavior for the nth social event, respectively
$z_{a_n}, z_{b_n}, z_{o_n}$	The latent sense in which a_n, b_n or o_n is used
$\mu_{z_{a_n},epa}$	The current expected E,P, or A value epa for the latent sense z_{a_n} for actor a_n. Similar entries exist for z_{b_n}, z_{o_n}
d_n	The difference between the actual deflection of the nth social event and the deflection expected from the fundamental
$\mu_{0,z_e,epa}$, $\sigma^2_{0,z_e,epa}$	The prior expected mean and variance for the EPA value epa for latent sense z_e of entity e.
$m_{0,z_e,epa}, k_{0,z_e,epa}$	Hyperparameters for $\mu_{0,z_e,epa}$
$v_{0,z_e,epa}, s_{0,z_e,epa}$	Hyperparameters for $\sigma^2_{0,z_e,epa}$
π_e	Distribution governing the likelihood of the different latent senses for entity e
p	Hyperparameter for π
α, ψ	Hyperparameters governing the prior likelihood of any identity or behavior, respectively
θ, ϕ	The estimated likelihood of any identity or behavior, respectively
I, B	The set of all identities or behaviors existent in all events in N, respectively
L	The assumed number of latent senses per identity and behavior
β	The expected scale of the distribution around d
Q	The parameter used to determine the likelihood of a behavior for a particular event
q	Dirichlet smoothing parameter for the conditional distributions $p(b\|o)$ and $p(b\|a)$

Table 2.2: Variables used in the description of the model

plate notation[9]. In this section, we introduce the model in accordance with its generative structure, working roughly from the top of Figure 2.1 to the bottom. Although the model is visually complex, we will show here that it is comprised of two rather straightforward pieces. First, the variables θ, ϕ, Q and their predecessors define a simple *language model* (Charniak, 1996), or a model which assigns probabilities to a sequence of words based on their distribution within a corpora of text. This language model governs the probabilities of drawing a particular actor/behavior/object combination for a social event. Second, the variables $\mu_0, \sigma_0^2, \pi, \mu_a, \mu_b, \mu_o, d, z$ and their predecessors define a sort of *Gaussian mixture model* (GMM) that uses ACT, which we will refer to as ACT-GMM. All variables we use, along with a brief description, are listed in Table 2.2. In reviewing the model, the reader may find Table 2.2 helpful in that it provides summaries of the mathematical constructs described here.

The model takes three forms of data as input. First, it accepts the set of social events N extracted from the dependency parser. Each social event in N consists of an actor a_n, a behavior b_n and an object o_n. For ease of notation, our discussion below assumes the n subscript on a, b and o is implicit. Second, model hyperparameters m_0 can be set to incorporate EPA profiles of entities appearing in N that also appear in the ACT dictionaries. Finally, the model accepts a change equation, used to calculate deflection. This equation is considered to be static and thus is

[9]For an introduction to plate notation, and to Bayesian modeling more generally, we refer the reader to the general texts from Gelman et al. (2013). We will here, out of necessity, assume some familiarity with such models

not updated in any fashion during model inference, nor is it explicitly referenced in Figure 2.1. The change equation we use is an average of the most recent female and male change equations as given by Interact as of December 30th, 2014.

2.4.1 Language model component

All entities are assumed to be drawn from a simple language model. The simplest feasible model would be to allow a, b and o to each be drawn from their own Categorical distributions, each of which has a Dirichlet prior. The use of a Dirichlet prior to "smooth" a multinomial or Categorical distribution is known as Laplace smoothing (Zhai and Lafferty, 2001). However, such a model would make poor use of both the data and the theory. With respect to the data, modeling the distributions of actors and objects separately ignores the fact that both draw from a common distribution over identities. Empirically, drawing from a single distribution over identities provides a significant improvement in the model's predictive capabilities. Additionally, ACT is often concerned with the behavior that connects two identities, and thus it makes more sense in the generative model to include an assumption that the behavior for an event is reliant on the actor and object. This assumption should exist in the language model, we believe, above and beyond similar assumptions wired into the ACT-GMM portion of the model.

In the language model, actors and objects are thus both assumed to be drawn from the same Categorical distribution θ, which defines a likelihood of the identity occurring in any given social event. Given an actor and an object, we then draw a behavior to connect them. We assume that the most likely behavior for this event will be influenced by the a and o selected, as well as the overall distribution of behaviors. This overall distribution of behaviors is encoded in the Categorical variable ϕ. The auxiliary variable Q, which is also Categorical, combines information in ϕ with Laplace smoothed estimates on the likelihood of b given a and o. We describe Q in more detail in the following section. Mathematically, these relationships can be expressed by the following:

$$
\begin{aligned}
\phi &\sim \text{Dirichlet}(\psi) \\
\theta &\sim \text{Dirichlet}(\alpha) \\
a &\sim \text{Categorical}(\theta) \\
o &\sim \text{Categorical}(\theta) \\
b &\sim \text{Categorical}(Q)
\end{aligned}
$$

2.4.2 ACT-GMM

Each identity and behavior in the dataset is assumed to have L possible EPA profiles in which it might be used within N, where L is set by the researcher and can be tuned empirically. Allowing for multiple EPA profiles for the same term is an important piece of our model, as the newspaper data we use is extracted from a variety of English-speaking cultures. Each culture may associate a unique EPA profile to a particular entity. We will refer to the different EPA profiles for a particular entity as its different *latent senses* in the sections below. The Categorical variable π governs the frequency with which each latent sense is expected to be used for each entity; p is a hyperparameter for π.

Each latent sense for each entity is associated with three values in μ_0 and σ_0^2; one for each dimension of the EPA profile for that latent sense for that entity. Here, and throughout the article, the 0 subscripts (e.g. on μ_0) are used to represent a variable that is a prior or a hyperparameter to the Bayesian network model. A particular entry in the vector μ_0, which we will refer to as $\mu_{0,z_{ib},epa}$ (ib stands for "identity or behavior") exists at the $3 * ib * z + epa$ location in μ_0. Here, z_{ib} is the index of the z_{ib}th latent sense for entity ib and epa is the index of the sentiment dimension. A similar indexing scheme is used for σ_0^2. Combined, these six mean and variance parameters determine the mean and variance of the three dimensions of the EPA profile for this particular latent sense z_{ib} of the entity ib.

All values in μ_0 are assumed to be drawn from a normal distribution governed by m_0, k_0 and σ_0^2, while σ_0^2 is assumed to be drawn from an Inverse Chi-squared distribution with parameters v_0, s_0. A key insight that is leveraged in our approach is that the values of m_0, as priors on the EPA profiles, can be used to set, or to "control", the EPA profiles for entities in the ACT dictionaries. For example, we might set the (static) value of $m_{0,1_{teacher},e} = 0.72$ to help ensure that the evaluative dimension of the first latent sense for the identity teacher is biased towards the "correct" value implied by the ACT dictionaries. More formally, we assume that the joint prior density for μ_0 and σ_0^2 follows a Normal Inverse Chi-squared distribution, which allows us to infer both values using Bayesian inference. This formulation is a common representation for Bayesian models where one wishes to infer both the standard deviation and the mean for a Normal distribution; as such we defer the reader to (Gelman et al., 2013, pp.76-68) for further details. Mathematically, our assumptions can be expressed as follows:

$$\begin{aligned}
\pi &\sim \text{Dirichlet}(p) \\
\mu_0 &\sim \text{N}(m_0, \frac{\sigma_0^2}{k_0}) \\
\sigma_0^2 &\sim \text{Inv}{-}\chi^2(v_0, s_0)
\end{aligned}$$

For each social event, each actor, behavior and object is associated with a particular latent sense z of its corresponding entity. Once z_a, z_b and z_o are drawn, we can obtain the entities' EPA profiles μ_a, μ_b, and μ_o (respectively) by sampling an EPA profile from the Normal distributions governed by the relevant entries in μ_0 and σ_0^2. Once these values have been drawn, we can obtain a deflection score for that event.

One could define the deflection for an event as a deterministic function. To do so, the values of μ_a, μ_b and μ_o would be combined to form the pre-event impression f. We could then provide a deterministic deflection score for the event by substituting these values into Equation (2.3). Instead, however, we treat deflection as a stochastic process whose mean is this expected deflection but that has some variance, β. We feel this assumption is more reasonable than the deterministic one in our particular case, as it accounts for context of this particular social event beyond what we can account for with our mixture model. For example, the lack of incorporation of information about settings implies an inherent randomness in the deflection measured by our model, thus justifying the assumption of stochasticity. The distribution of deflection is assumed to be Laplacian, which makes model inference easier while still retaining the desired sociotheo-

retic meaning of deflection as a distance metric. Mathematically, our assumptions can be stated as follows:

$$z_a \sim \text{Categorical}(\pi) \qquad z_b \sim \text{Categorical}(\pi) \qquad z_o \sim \text{Categorical}(\pi)$$
$$\mu_a \sim \text{N}(\mu_{0,z_a}, \sigma^2_{0,z_a}) \qquad \mu_b \sim \text{N}(\mu_{0,z_b}, \sigma^2_{0,z_b}) \qquad \mu_o \sim \text{N}(\mu_{0,z_o}, \sigma^2_{0,z_o})$$
$$d \sim \text{Laplace}(\sum_i^9 (f_i - M^T_{i*}G(f))^2, \beta) \quad \text{where} \quad f = [\mu_{a_e}, \mu_{a_p}, \mu_{a_a}, \mu_{b_e}, \mu_{b_p}, \mu_{b_a}, \mu_{o_e}, \mu_{o_p}, \mu_{o_a}]$$

2.4.3 Summary

Having fully defined our model, a useful summarization can now be provided by giving the formal generative process required by the model. To generate a new social event, the following process is carried out:

1. Draw an actor and an object; $a \sim \text{Cat}(\theta) \qquad o \sim \text{Cat}(\theta)$
2. Draw a behavior; $b \sim \text{Cat}(Q)$
3. Draw a latent sense for a, b and o; $z_a \sim \text{Cat}(\pi_a) \qquad z_b \sim \text{Cat}(\pi_b) \qquad z_o \sim \text{Cat}(\pi_o)$
4. Draw EPA profiles for a, b, and o
   - $\mu_{a,e} \sim \text{N}(\mu_{0,z_a,e}; \sigma^2_{0,z_a,e}) \quad \mu_{a,p} \sim \text{N}(\mu_{0,z_a,p}; \sigma^2_{0,z_a,p}) \quad \mu_{a,a} \sim \text{N}(\mu_{0,z_a,a}; \sigma^2_{0,z_a,a})$
   - $\mu_{b,e} \sim \text{N}(\mu_{0,z_b,e}; \sigma^2_{0,z_b,e}) \quad \mu_{b,p} \sim \text{N}(\mu_{0,z_b,p}; \sigma^2_{0,z_b,p}) \quad \mu_{b,a} \sim \text{N}(\mu_{0,z_b,a}; \sigma^2_{0,z_b,a})$
   - $\mu_{o,e} \sim \text{N}(\mu_{0,z_o,e}; \sigma^2_{0,z_o,e}) \quad \mu_{o,p} \sim \text{N}(\mu_{0,z_o,p}; \sigma^2_{0,z_o,p}) \quad \mu_{o,a} \sim \text{N}(\mu_{0,z_o,a}, \sigma^2_{0,z_o,a})$
5. Draw a deflection score for the event
   - $d \sim \text{Laplace}(\sum_i^9 (f_i - M^T_{i*}G(f))^2, \beta) \quad \text{where} \quad f = [\mu_{a_e}, \mu_{a_p}, \mu_{a_a}, \mu_{b_e}, \mu_{b_p}, \mu_{b_a}, \mu_{o_e}, \mu_{o_p}, \mu_{o_a}]$

The described process helps to explain how a new social event might be "generated" by the Bayesian network model described here, but also provides insight into how the model determines the likelihood of an event it is given. The likelihood of a particular event is a function of likelihood of the actor and object's identities overall (1.), the "semantic likelihood" of the behavior given these identities (2.), and the "affective likelihood" of the social event as a whole (3.-5.).

2.5 Model inference

Model inference is completed in two steps. First, we determine Maximum *a posteriori* (MAP) for the parameters of the language model, as they are straightforward enough to determine in closed form. Second, we use the Stochastic EM (Tregouet et al., 2004) (Bishop and others, 2006, p. 439) algorithm displayed in Algorithm 1 to draw inferences for parameters in the GMM portion of the model. Note that Algorithm 1 references several equations that will be introduced later in this section. In the Expectation ("E") step, we use Gibbs sampling to draw expected values for z_a, z_b and z_o and for μ_a, μ_b and μ_0 for all social events. In the Maximization ("M") step, we then update σ_0, μ_0 and π with their MAP estimates. Note that we do not explicitly sample d, as the value of this stochastic process is not of particular interest to us in the present work.

Below, we first derive the MAP estimate for the language model. We then derive the Gibbs sampling equations for all z and all μ and finally the MAP estimates for σ_0, μ_0 and π. In doing

Algorithm 1: Model inference for ACT-GMM portion using Stochastic EM

Initialize all μ_0, σ_0^2, π
for $i = 0$ **to** $N\_ITERATIONS$ **do**
    E step:
        Sample all z using Gibbs sampling and Equation (2.7)
        Sample all μ using Gibbs sampling and Equation (2.8)
    M step:
        Update all σ_0 using MAP estimate of Equation (2.10)
        Update all μ_0 using Equation (2.12)
        Update all π using Equation (2.13)

so, we introduce three additional pieces of notation. First, let μ_* represent all nine fundamental values drawn for an event. Second, let μ_*/x represent μ_* where all values of all elements are known except for x. Finally, we define Ω as the set of all parameters, and Ω_{-x} as the set of all parameters besides x.

2.5.1 MAP estimates for language model

MAP estimation for the language model portion of the model is relatively straightforward. The variables of interest are θ and Q_n. The distribution for θ is given in Equation (2.4), where $n(a_i)$ is a function that represents the number of times the identity i appeared as an actor in N and $n(o_i)$ the number of times i appeared as an object. The Dirichlet distribution is a well-known conjugate of the Categorical distribution, and we thus do not re-derive the posterior distribution here. Note, however, that we follow the notational convenience of absorbing the minus one in the second line of Equation 2.4 into the Dirichlet hyperparameter in all following statements about the posterior distribution and MAP estimates of the Categorical distribution.

$$\begin{aligned} p(\theta) &= p(\theta|\alpha) * \prod_{i=1}^{|I|} p(a_i|\theta)p(b_i|\theta) \\ &\propto \prod_{i=1}^{|I|} \theta_i^{n(a_i)+n(o_i)+\alpha_i-1} \\ &\sim \text{Dirichlet}(n(a_i) + n(o_i) + \alpha_i) \end{aligned} \tag{2.4}$$

Given $p(\theta)$ is distributed as in Equation (2.4), the MAP estimate for the posterior distribution of θ is given by Equation (2.5). The estimate is simply a normalized function of the number of times an identity appears plus the "pseudo-counts" from the Dirichlet prior α.

$$\hat{\theta} = \arg\max_\theta p(\theta) = \frac{n(a_i) + n(o_i) + \alpha}{\sum_{i \in I} n(a_i) + n(o_i) + \alpha} \tag{2.5}$$

The MAP estimator for Q is given in Equation (2.6). Note that because Q depends on the actor and objects for each event, there are actually $|N|$ values of Q. We will discuss the derivation for a particular entry of Q, Q_n here, as the derivation is the same for all events. The distributions $p(b_n|a_n)$ and $p(b_n|o_n)$ are Categorical distributions that give the conditional likelihood of the

behavior b_n given a_n and o_n, respectively. To ensure that these values are never zero, we introduce a smoothing parameter q resulting in the distributions $p(b_n|a_n, q)$ and $p(b_n|o_n, q)$, respectively. Introducing the smoothing parameter q is equivalent to inserting an auxiliary variable for both $p(b_n|a_n)$ and $p(b_n|o_n)$ and putting a Dirichlet prior over each with the hyperparameter q. As the introduction of this variable would unnecessarily complicate notation, we do not use it here. The likelihood of any particular behavior as derived from the MAP estimate is thus simply the product of three Laplace smoothed Categorical variables, ϕ (smoothed by ψ), $p(b_n|a_n)$ and $p(b_n|o_n)$, both smoothed by the constant q. The distribution of b_n is thus Categorical with Q as the parameter.

$$\begin{aligned}\hat{Q_n} = \arg\max_{Q_n} p(Q_n) &= \arg\max_{Q_n} p(b_n|a_n, q) * p(b_n|o_n, q) * p(\phi|\psi) \prod_i^N p(b_{n,i}|\phi) \\ &= \frac{n(b_i|a) + q}{\sum_{b_i \in B} n(b_i|a) + q} * \frac{n(b_i|o) + q}{\sum_{b_i \in B} n(b_i|o) + q} * \frac{n(b_i) + \psi}{\sum_{b_i \in B} n(b_i) + \psi} \end{aligned} \quad (2.6)$$

2.5.2 "E" Step for ACT-GMM

For each document, we must draw z_a, z_b and z_o and all nine values for the fundamental, three each for μ_a,μ_b and μ_o μ. Because the sampling proceedure is analogous for all entities in a particular event and are the same for each event, we will focus here only on the agent for one specific event n.

Sampling z

The conditional probability that the variable z_{a_n} is equal to the latent sense t is specified in Equation (2.7):

$$p(z = t|\Omega_{-z}) = p(z = t|\pi) \prod_i^{[e,p,a]} p(\mu_{a_i}|\mu_{0,t,i}, \sigma^2_{0,t,i}) \quad (2.7)$$

Sampling from this conditional distribution is straightforward, as both the first and second terms of the probability function are easy to compute. The first term is simply the likelihood of latent sense t as given by the current value of π. The second term can be obtained by evaluating the likelihood of $\mu_{a,e}$,$\mu_{a,p}$ and $\mu_{a,a}$ relative to their expected distribution given the current state of μ_0 and σ_0. These three values are multiplied together to generate a likelihood for μ_a as a whole. Multiplying the result of this process by the first piece of the probability function, we can then normalize over all possible values of z and draw a new latent sense for the actor in this event from this Categorical distribution.

Sampling μ

The conditional distribution of $\mu_{z_{a_n},e}$, the value for the evaluative dimension of the actor's EPA profile for event n and latent sense z_{a_n}, is given below in Equation (2.8). Representation of the potency and activity dimensions are analogous, so we focus only on the evaluative dimension here. Also, we shorten z_{a_n} to z ease notation.

$$p(\mu_{z,e}|\Omega_{-\mu_{z,e}}) = p(\mu_{z,e}|\mu_{0,z,e}, \sigma^2_{0,z,e})p(d|\mu_*/\mu_{z,e}; \beta) \tag{2.8}$$

To infer the conditional distribution of $\mu_{z,e}$, Equation (2.8) shows we simply need to understand the prior distribution of $\mu_{z,e}$ and the distribution of d given all values except that of $\mu_{0,z,e}$. From the section above, we know that $p(\mu_{z,e}|\mu_{0,z,e}, \sigma^2_{0,z,e}) \sim \mathrm{N}(\mu_{0,z,e}, \sigma^2_{0,z,e})$. Thus, we are left with interpreting the distribution of $p(d|\mu_*/\mu_{z,e}; \beta)$. It can be shown, rather unexpectedly that evaluating the distribution of d given all values except $\mu_{z,e}$ results in a distribution which is normally distributed on $\mu_{\mathbf{z,e}}$ with a known mean and variance.

The proof is shown below; the derivation follows from the fact stated in Section 2.2.1 that the deflection score with one unknown variable is a quadratic in that variable. By completing the square and dropping constant terms that do not inform the conditional distribution for $\mu_{z,e}$, we are left with a function that defines a Normal distribution on $\mu_{z,e}$ with the given parameters.

$$\begin{aligned}
p(d|\mu_*/\mu_{z,e}; \beta) &\propto \exp(-\frac{|d - \sum_i^9 (f_i - MG(f_i))^2|}{\beta}) \\
&= \exp(-\frac{|d - (c_0\mu_{z,e}^2 + c_1\mu_{z,e} + c2)|}{\beta}) \\
&\propto \exp(-\frac{|(c_0\mu_{z,e}^2 + c_1\mu_{z,e})|}{\beta}) \\
&= \exp(-\frac{|c_0|(\mu_{z,e} + \frac{c_1}{2c_0})^2}{\beta}) \\
&= \exp(-\frac{(\mu_{z,e} + \frac{c_1}{2c_0})^2}{\frac{\beta}{|c_0|}}) \\
&\propto \mathbf{N}_{\mu_{\mathbf{z,e}}}(-\frac{c_1}{2c_0}, \frac{\beta}{2|c_0|})
\end{aligned} \tag{2.9}$$

There are two important points to note in the derivation shown in Equation (2.9). First, the result relies on the fact that there are no social events in which the same identity appears more than once. If this were to be the case, the equation would no longer be quadratic in $\mu_{z,e}$. Second, and perhaps more importantly, is that the resulting distribution is centered at the value of $\mu_{z,e}$ which minimizes the deflection of the social event given all other fundamental meanings as estimated by Maximum Likelihood Estimation (Heise, 2007, ch. 8). Though this result fits our intuition, we do not believe that this was an obvious outcome given the initial distribution.

Thus, when updating $\mu_{z,e}$ we are drawing from a product of two normal distributions. One of these distributions is centered at the current expected value of $\mu_{z,e}$ as given by $\mu_{0,z,e}$ and $\sigma^2_{0,z,e}$. The second distribution, usefully, is centered at the value which will minimize deflection for the current event given all other values in the pre-event fundamental vector. It is well-known that the product of two normals is proportional to a new normal distribution[10], and thus we can sample a new value for $\mu_{z,e}$ from this new distribution, which has a mean of $\frac{\mu_{0,z,e} * \frac{\beta}{2|c_0|} - \frac{c_1}{2c_0} * \sigma^2_{0,z,e}}{\sigma^2_{0,z,e} * \frac{\beta}{2|c_0|}}$ and a

[10] for a formal proof, see (Bromiley, 2013)

variance of $\frac{\sigma^2_{0,z,e} * \frac{\beta}{2|c_0|}}{\sigma^2_{0,z,e} + \frac{\beta}{2|c_0|}}$. From this sampling distribution, it is clear that the new value is informed by both the prior information from μ_0 and σ^2_0 and information from the current event.

2.5.3 "M" Step for ACT-GMM

MAP estimates for μ_0, σ^2_0

Because all updates in μ_0 and σ^2_0 are analogous, we will consider the conditional distribution of the evaluative dimension of a particular latent sense z_i of a single identity i. To ease notation, we will refer to the relevant entry in μ_0, which is $\mu_{0,z_i,e}$, as simply μ_0, and the relevant entry in σ^2_0, which is $\sigma^2_{0,z_i,e}$, as simply σ^2_0. A similar shortening of notation will be applied to the four relevant hyperparameters $m_{0,z_i,e}$, $k_{0,z_i,e}$, $s_{0,z_i,e}$ and $v_{0,z_i,e}$. Let us also define the set S, which consists of all events in which the latent sense z_i of the identity i is used in the current iteration of the inference algorithm. Formally, $S = \{n \in N : (a_n = i \,\&\, z_{a_n} = z_i) \,|\, (o_n = i \,\&\, z_{o_n} = z_i)\}$. The variable S is introduced as we need not worry about events outside of it; they will be irrelevant in evaluating the distribution of $\mu_{0,z_i,e}$.

The derivation for the MAP estimation of σ^2_0 can be easily obtained from its well-known posterior distribution, shown in Equation (2.10). In Equation (2.10), $\bar{s^2}$ is the sample variance deviation (that is, $\sum_n^S (\mu_n - \mu_0)^2$) and $\bar{\mu} = \frac{\sum_n^S \mu_n}{|S|}$.

$$\begin{aligned} p(\sigma^2_0 | \Omega_{-\sigma^2_0}) &= p(\sigma^2_0 | v_0, s_0) \prod_n^S p(\mu_n | \mu_0, \sigma^2_0) \\ &= \text{Inv}{-}\chi^2(v_0 + |S|, v_0 s_0 + (|S| - 1) * \bar{s^2} + \frac{k_0 * |S|}{k_0 + |S|}(\bar{\mu} - \mu_0)^2) \end{aligned} \tag{2.10}$$

The expected value of this posterior distribution is then the MAP estimate of Equation (2.10). The MAP estimate is $\frac{xy}{x-2}$, where x is the location (first parameter) of the $\text{Inv}{-}\chi^2$ in Equation (2.10) and y is the scale (second parameter) of this distribution.

The distribution of μ_0 also consists of two parts, a straightforward prior and a posterior component that is the product across the events in S. This is shown in Equation (2.11):

$$p(\mu_0 | \Omega_{old}) = p(\mu_0 | m_0, \sigma^2_0, k_0) \prod_n^S p(\mu_n | \mu_0, \sigma^2_0) \tag{2.11}$$

The MAP estimate of a normal distribution in this form is well known, so we simply provide the resulting estimate in Equation (2.12). Note that σ^2_0, as used in Equation (2.12), represents the "new" version of σ^2_0 from Equation (2.10).

$$\hat{\mu}_0 = \frac{\frac{k_0}{\sigma^2_0} m_0 + \frac{|S|}{\sigma^2_0} \bar{\mu}_n}{\frac{k_0}{\sigma^2_0} + \frac{|S|}{\sigma^2_0}} \tag{2.12}$$

MAP estimate of π

The MAP estimate of π reduces to a new Categorical distribution where the likelihood of each latent sense is the number of times this latent sense is "used" in the E step plus the "pseudo-counts" from the Dirichlet prior p. The derivation of this MAP estimate is, as we have mentioned,

Variable	Initialization value/method
β, α, ψ, q	1
v_0	2
s_0	.1
p	3
m_0	value from ACT dictionary, random otherwise
k_0	50 if from ACT dictionary, 10 or 1 otherwise. See text
L	varied for parameter tuning
π, μ_0, σ_0^2	drawn from Prior

Table 2.3: Model initialization details

straightforward given previous, well-known results in the literature. Equation (2.13) gives the distribution of the new value of π.

$$\hat{\pi} = \frac{n(z_t) + p_t}{\sum_s (n(z_s) + p_s)} \tag{2.13}$$

2.5.4 Initializing the Model

All that remains to be introduced with respect to model inference is how parameters are initialized. Table 2.3 details the initialization of all parameters aside from Q, d, z and μ. An initialization of Q is unnecessary, as we simply compute it's MAP estimate once. The value of d is not of interest and does not affect the estimation of other parameters, thus initialization is unnecessary. The parameters z and μ are sampled before they are used, so also need not be initialized. In this section, beginning with the hyperparameters, we provide more details about the meaning of the statements in Table 2.3.

Hyperparameter tuning can have important implications on model performance, perhaps especially so in cases where we are dealing with language data (Wallach et al., 2009). However, hyperparameters also reflect one's prior expectations, and thus we attempt here to balance a search for optimal parameters between our prior expectations and model performance. Further, given the number of hyperparameters for the model, we chose to only use heuristic searches to explore the parameter space. Thus, we set $\beta, q, p, \alpha, \psi, s_0$ and v_0 to specific values based on the parameter settings which maximized performance on a single fold of the data and set m_0 and k_0 via a combination of informal model testing and heuristic methods.

For β, we assume that a reasonable prior for the variance of the deflection score around its ACT-implied value is 1. The variable q is set to 1, as testing on development data using a variety of language models suggested that this minimal pseudo-count led to the strongest model. We also set α and ψ to 1 based solely on results from testing, as these values performed better than other tested values of 3, 50 and 1000. We similarly set s_0 to .1 as opposed to other tested values of .5 and 1, and p to 3 as opposed to other tested values of 50, 400 and 1000. Finally, we set v_0, which can be thought of as our confidence in the prior s_0, to 2, which reflects a low level of confidence in the value of s_0. Setting the value to 1 caused high instability in the MAP estimates for σ_0^2; the value of 2 was the smallest at which they showed stability.

We use the hyperparameter m_0 to encode our prior belief of the EPA profile for each latent sense of each entity. We make the assumption that one of the latent senses is drawn from an American cultural perspective, as many of the media outlets within our dataset are based in America. Hence, we use data from the ACT website that is the best representation avaiable of this perspective. This data comes from a dictionary of 500 identities and 500 behaviors coded with EPA profiles from surveys in 2002-3 of undergraduate students at a large, public, American institution (Francis and Heise, 2006). These data have been used in a variety of ACT studies since it was introduced, and are used as the default values for the computer program Interact (Heise, 2010a), from which much ACT research is derived.

For entities ib that are in both the survey data and our set of social events, we initialize values of m_0 for its zeroth sense (i.e. $m_{0,0_{ib},e}, m_{0,0_{ib},p}, m_{0,0_{ib},a}$) to the values from the survey data. We then heuristically set the rest of the values of m_0 using an iterative algorithm for the zeroth sense of all entities that are not in the ACT dictionary. The algorithm takes as input the set of already known values in m_0 and uses the fact that the standard mathematical model of ACT can be used to solve for the EPA profile of the third entity in an event if the EPA profiles of the other two entities are already known (Heise, 2007, ch. 8). On the first round of the iterative algorithm, we extract all events where values for m_0 for two of the three entities are known from the ACT dictionary. We then compute the optimal EPA profile for the third entity in each of these events. We then take the average EPA profile for each entity we can obtain at least one EPA profile from in this process and use these values to initialize $m_{0,0_{ib},e}, m_{0,0_{ib},p}$ and $m_{0,0_{ib},a}$ for each of these entities.

Once we have set these values, we can iterate through N again, treating both the original m_0 values from the ACT dictionary and the new values from the first iteration of the algorithm as known. This iterative process continues until we can initialize m_0 for all entities. If we reach a point where no new information can be gleaned from the process above, we select one random event and set the EPA profile of one of the entities using uniform random values in the range $[-4.3, 4.3]$. This allows the algorithm to continue learning. In practice, the algorithm finishes in around two iterations, only having to set a random score zero or one times for entities that appear in the training data. For terms appearing in held-out test data but not in the training data, the appropriate entries for the zeroth sense in m_0 are initialized to uniform random values.

The zeroth sense for each entity thus represents some instantiation, real or heuristically defined, of the EPA value of that entity from an American cultural standpoint. Values in m_0 for all other latent senses of all entities (including those found in the ACT dictionary) are set using uniform random values on the interval $[-4.3, 4.3]$. While future work may attempt to be smarter in how these parameters are set, we currently use random values to insinuate no prior knowledge of the EPA profiles of other cultural groups whose perceptions may exist in the data. This lack of prior knowledge is reiterated with the initialization of k_0, which can be thought of as the number of observations that we associate with m_0 as a prior for μ_0. For the zeroth latent sense for entities in the ACT dictionary, we set k_0 to 50 (the number of respondents in the survey data). For the zeroth latent sense of entities initialized in the iterative algorithm, we set k_0 to 10. For all other latent senses of all entities, $k_0 = 1$.

Once we have initialized all hyperparameters, all that remains is to initialize are π, μ_0 and σ_0^2. We do so by drawing π, μ_0 and σ_0^2 from their respective prior distributions. This completes model initialization.

UNIGRAM	$p(b\|a,o) = p(b, s = 1)$
BIGRAM	$p(b\|a,o) = p(b, s = 1)p(b\|o, s = 1)p(b\|a, s = 1)$
ACTONLY	$p(b\|a,o) = p(b\|\phi, a, o, q)p(b\|a, o, d = 0, m_0)$
NOACT	$p(b\|a,o) = p(b\|\phi, a, o, q)\mathbf{E}_{z,\mu}[p(z\|\pi)p(d = 0\|\mu)p(\mu\|\mu_0, \sigma_0^2)]$
FULLMODEL	$p(b\|a,o) = p(b\|\phi, a, o, q)\mathbf{E}_{z,\mu}[p(z\|\pi)p(d = 0\|\mu)p(\mu\|\mu_0, \sigma_0^2)]$

Table 2.4: Predictive distributions for the four baseline models and the full model

2.6 Approach to Model Evaluation

The model we present is a combination of three well-established methodological approaches-language modeling, Gaussian mixture models and ACT. While we have confidence that each component can extract useful information from the data, our extensions of current ACT methodology and the novel way in which we combine techniques requires a careful study of the extent to which our efforts produce parameter estimates that are truly representative of information within the data. To evaluate the quality of the estimates generated by our model, we use 10-fold cross-validation. In k-fold cross validation, the data is split into k "folds". We use $k - 1$ folds as "training data" to train the model and carry out a prediction task on the "left out" test data. This process is repeated k times, leaving out a different chunk of the data, and then results on the prediction task across all folds are averaged. Here, we use perhaps the most common prediction task in the ACT literature, that of predicting the behavior between an actor and an object. That is, for a given left out event n, we give the trained model the actor and object in n and then attempt to predict the behavior.

In establishing the quality of our model's predictions, we can have more confidence that parameter estimates accurately represent processes within the social event data. Importantly, such an understanding requires some baseline for comparison. The four baseline models we compare our full model's predictions to range from simple language models to more complex structural ablations of our full model. While only some of these models can actually help to infer EPA profiles of entities, they are important in giving us a sense of how "easy" the prediction task we are addressing is. When a simple baseline can predict the data perfectly, a complex model like the one we propose is more likely to learn patterns in the data that are largely noise, and thus post-hoc interpretations of parameter estimates may suffer.

To evaluate the success of each model on the prediction task, we compute the *log perplexity* of the model in determining the correct behavior across all test events. Log perplexity, or simply perplexity as it is often written, a measure of accuracy typically used in explorations of the predictive abilities of NLP models. This value is defined in Equation (2.14) below. In the equation, let TD be the set of held out data used for testing for a single fold. The log perplexity, averaged across all test events in all folds, gives us a sense as to how much weight in the model's predictive distribution that the model places on the correct behavior. The value $2^{log(perpl(TD))}$ can be thought of as the number of behaviors the model feels are equally as likely as the true answer. So, if $log(perpl(TD)) = 1$, the model would, on average, be "flipping a coin" between the correct answer and one other answer. If $log(perpl(TD)) = 4$, the model would be rolling a 16-sided die. Note that the metric is a measure of the extent to which a particular model is "confused" by the data, and thus a lower score represents a better model.

$$log(perpl(TD)) = \frac{-\sum_{n \in TD} log(p(b_n|a_n, o_n))}{|TD|} \tag{2.14}$$

All models we test and the predictive distributions they use to determine the likelihood of each behavior for a given test event are shown in Table 2.4. The first two models are simple, Laplace smoothed language models with a smoothing parameter $s = 1$. The first predicts the likelihood of a behavior by simply determining the likelihood of the single-word behavior label, or the behavior *unigram*, $p(b, s = 1)$. We call this model the UNIGRAM model. The second model uses the conditional likelihood of the behavior given the actor and the object independently, as well as the likelihood of the behavior itself. This is exactly the language model used in our full Bayesian model. Drawing from the language modeling literature, this model is termed the BIGRAM model, meaning that we draw information for the behavior from its distribution, its distribution conditioned on the actor and its distribution conditioned on the object.

Combined, these two baselines show how well events can be predicted by considering only the semantic relationships between identities and behaviors. Note that this semantic information, particularly in the BIGRAM model, will implicitly capture a significant amount of affective meaning as well - just because we are not explicitly modeling affecting meaning does not mean it isn't capture in semantic relationships within the text. We should therefore expect that these semantic models, which derive likelihoods from only connections between words, are strong predictive models. Adding in the ACT component of the model may or may not help in a predictive sense, but will serve the vital purpose of helping us to understand *why* these semantic relationships are occurring.

The third baseline we use removes the GMM portion of the full model, replacing it with what is essentially the pure prediction model of current ACT methodologies. In this ACTONLY model, we use the values of m_0 initialized by the iterative algorithm described above to determine the sentiment for each entity. These values are therefore only roughly informed by the data, capturing only the heuristic information from the initialization algorithm. The ACTONLY model then uses these heuristically set EPA profiles in combination with the language model and the base mathematics of ACT to make predictions, no further statistical optimization is performed. Under the assumption that the actual deflection of each social event is zero, i.e. $d = 0 \, \forall \, n$, and that $\beta = 1$, the prediction reduces to a form of the probability model for deflection related to that proposed by Hoey et al. (2013a), as shown in Equation (2.15).

$$\begin{aligned} p(b|a, o) &= p(b|\phi, a, o, q) * p(b|a, o; m_0, d = 0, \beta = 1) \\ &= p(b|\phi, a, o, q) * \exp(\sum_{i}^{9}(f_i - MG(f_i))^2) \\ \text{Where} \quad f &= [m_{0,a_e} \;\; m_{0,a_p} \;\; m_{0,a_a} \;\; m_{0,b_e} \;\; m_{0,b_p} \;\; m_{0,b_a} \;\; m_{0,o_e} \;\; m_{0,o_p} \;\; m_{0,o_a}] \end{aligned} \tag{2.15}$$

The final baseline we use is one that effectively removes ACT from the full model by randomizing the change equation matrix and removing any information from the ACT dictionaries in the priors. This model is labeled the NOACT model in our results. While we expect performance of this model to be comparable to the full model, it loses a significant amount of value in

qualitative analysis of results. Finally, we of course train our FULL model. Note that we run both of these models assuming a variety of values for L in order to understand how they perform with different numbers of assumed latent senses.

For both the NOACT and FULL models, we run Algorithm 1 for 200 iterations. Parameter estimates used in the prediction task are extracted from the final iteration of the algorithm. Once parameter estimates have been obtained, we also must account for the fact that in both models, the likelihood of a particular behavior for a test event is determined by averaging over all possible values of z_a, z_b and z_o and all values of μ_a, μ_b and μ_o. That is, we must compute $\mathbf{E}_{z,\mu}[p(d = 0|\mu)p(\mu|\mu_{0,z}, \sigma^2_{0,z})p(z|\pi)]$, which when expanded becomes $\sum_z \int_\mu p(d = 0|\mu)p(\mu|\mu_{0,z}, \sigma^2_{0,z})p(z|\pi)$. Note that we here condense all z and μ values for the actor, behavior and object into a single term to simplify notation.

We choose to estimate this expectation using Gibbs sampling. To do so, we can simply draw z_a, z_b and z_o and then μ_a, μ_b and μ_o. After making $|S|$ such draws and computing the value of $p_s(d = 0|\mu_s)$ for each draw s, we then can get an estimate of the likelihood of any actor/behavior/object combination. Formally, we use the fact that $\mathbf{E}_{z,\mu}[p(d = 0|\mu)p(\mu|\mu_{0,z}, \sigma^2_{0,z})p(z|\pi)] \approx \frac{\sum_s^{|S|} p_s(d=0|\mu_s)}{|S|}$. We use 50 Gibbs samples each time we compute the expectation.

2.7 Results

2.7.1 Prediction Task

Figure 2.2 displays results for log perplexity (recall, the equation for log perplexity was given in Equation (2.14)) for the four baseline models we used, as well as the full model. On the y-axis, the average perplexity across folds is given. Note that this value is only computed across nine of the ten folds; we ignore results from the fold we use to determine hyperparameter values. The x-axis represents the numbers of latent senses used in the model. For models that do not use multiple senses (the UNIGRAM, ACTONLY and BIGRAM models), Figure 2.2 shows two vertical bands which represent the upper and lower limits of the 95% bootstrapped confidence intervals. For both the NOACT and FULL models, we run iterations with 1, 2, 3, 5 and 7 latent senses and present confidence intervals for models evaluated at each.

Figure 2.2 shows that the worst performing model was the UNIGRAM model. As this is the simplest possible approach to behavior prediction, this is not surprising. However, the poor performance of the unigram model relative to the others is important in that gives us confidence that the prediction problem is non-trivial. The ACTONLY model, which implements the basic ACT model, improves our ability to predict behavior by almost an order of magnitude over the baseline UNIGRAM approach. Similarly, the the FULL model performs significantly better than the NOACT model when the number of latent senses is low.

More specifically, we see that affective information encoded in the ACT model improves perplexity by 16% with only one latent sense and by 3% when the number of latent senses per entity is assumed to be two. These gains, while modest are statistically significant - 95% CIs do not overlap at all in either case. However, Figure 2.2 also shows that as the number of latent senses increases, the difference in predictive power between these two models begins to decrease. Specifically, as the number of assumed senses extends beyond three, model performance on the prediction task becomes virtually indistinguishable. This shows that as the number

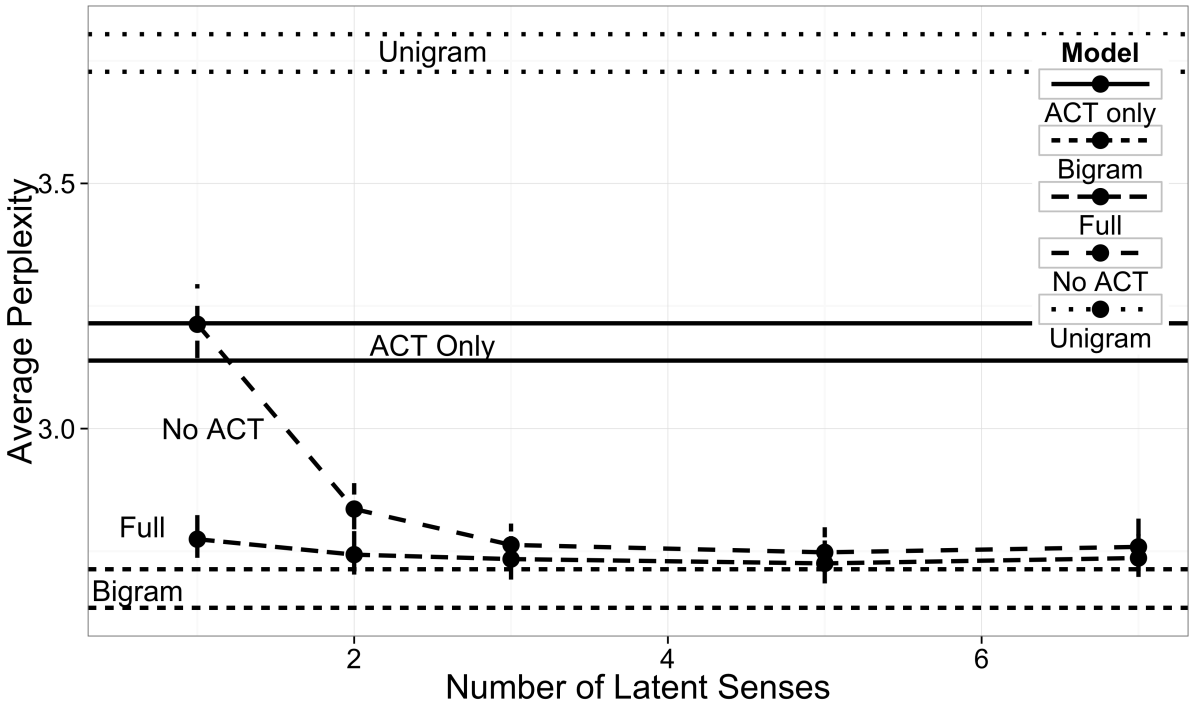


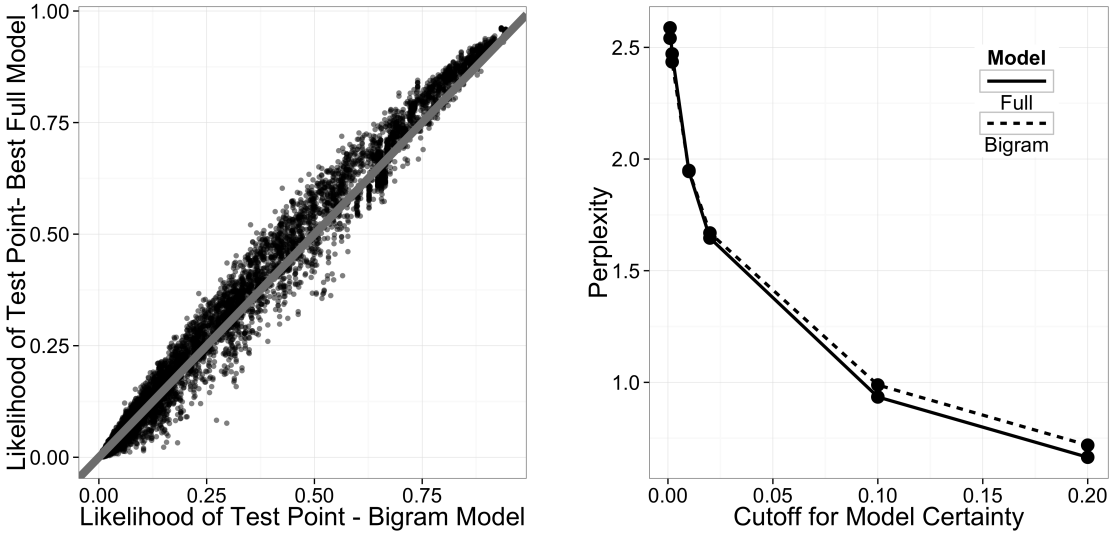
Figure 2.2: Average Perplexity for the four baseline models and the full model for varying numbers of topics. Where models do not use topics, we present two bands, the top and bottom of the 95% bootstrap CIs. For the FULL and NOACT models, 95% bootstrap CIs are also shown for the different values of L at which the models were trained.

of latent senses increases, the GMM portion of both the NOACT and FULL models is able to find parameter values that can reliably inform us of future events.

We therefore see that when the number of parameters in the model is low, the theory of ACT provides important guidance for how any assumed affective meaning is structured. As the number of free parameters in the model grows, however, the need for a theoretically driven model of affect decreases in order to accurately predict the data - the model is able to fit the data well in either case because it has enough parameters to "make up" for the lack of theoretically driven priors. Predictive accuracy in this case, however, is still sacrificed for the use of the resulting parameters. As noted in the sections above, only results for the FULL model are useful in interpreting EPA profiles of entities, as only the full model allows us to begin from a baseline of intuitive EPA values for at least a subset of the data. In other words, while the NOACT model assumes the existence of affective constraints, it is essentially free to "make up" its own cultural norms about the form of those constraints. Only in the FULL model, where affective constraints have been painstakingly estimated via decades of survey data, are the affective constraints estimated by the model likely to match our culturally-normed intuitions.

Finally, Figure 2.2 shows that neither the FULL nor the NOACT models perform as well as the BIGRAM model. This indicates that the ACT-GMM portion of the model actually *decreases* the predictive performance of the bigram language model on the training data. This stems from two factors. First, as mentioned above, semantic information from the BIGRAM model retains a large

amount of the affective meanings which may drive these semantic connections. Again, though, it is only with the FULL model that we are able to better understand these affective relationships. Second, because the BIGRAM model is a probability model over only information in the training data, it does not need to "consider" information from the ACT dictionary, information which does not always confirm that provided in the dataset of social events.



(a) (b)

Figure 2.3: a) Comparison of the predictions for the best full model and the bigram model for all test points - a dashed line is drawn where x=y for clarity; b) Comparison of average perplexity of the full and bigram models for different likelihood cutoffs

Regardless of this difference in the function that these two models maximize, predictive ability on the testing data is the best tool available to understand how well parameter estimates represent the social event data. To better understand how the ACT-GMM portion of the model affects predictions from the language model, Figure 2.3a plots the likelihood given to the correct behavior for each test point across the nine folds. Each point on the graph represents a single test point. The y-value of a point represents the probability of the actual behavior in the event as evaluated by the best FULL model (where $L = 5$). The x-value provides the same probability, except from the BIGRAM model. Points that fall above the grey diagonal line are those in which the full model assigned a higher probability to the actual behavior for the test point, while points below the line represent those where the bigram model outperformed the full model.

Figure 2.3a is constructed with alpha blending, so darker areas of the plot represent areas where more test points fell. Most test points fall near the two tails of the likelihood- that is, either both models believed the test point to be highly unlikely or highly likely. As we see, the bigram model performed significantly better at the lower tail. Thus, the bigram model put more weight in the posterior predictive distribution on the correct behavior in cases where both models believed the true behavior to be highly unlikely. In contrast, the full model performs much better on test points that both models believed were more likely than chance (any value of less than approximately .011 represents a point for which both models would do worse than a random

guess). Consequently, we can have some confidence

Figure 2.3b emphasizes this point. On the y-axis, average perplexity is given for the best full model (solid line) and the bigram model (dashed line). On the x-axis, we provide a cutoff value for test points. As the cutoff increases, we remove test events where both models put less probability on the correct behavior than the cutoff value. Thus, for a cutoff of .02, we remove all test events where both models put less than 2% of the weight in their respective posterior predictive distributions on the correct behavior. Figure 2.3b shows that full model performance is slightly better than the simple bigram model on data points for which both models believe the actual behavior to be relatively likely.

Combined, Figure 2.3a and Figure 2.3b suggest that information from the ACT-GMM portion of the full model aids the language model portion in predicting already likely behaviors, and detracts in cases where the behavior is already unlikely. This suggests that the model that we have developed struggles when given noisier data. It also suggests, however, that the affective meaning the model uses for prediction may provide important information in a predictive sense beyond what one can derive from pure semantics. This observation shows that a better pipeline of event extraction and future iterations of the model may be provide an important new avenue of predictive modeling of text as well.

As stated, however, the goal of the present work is not to predict behavior, but rather to infer affective meaning. Results in this section were therefore intended to show that the full model learns parameter estimates indicative of the data as proven by the model's ability to predict events it has not seen. To this end, we observe that the full model performs significantly better than all baselines except the bigram language model. We noted one reason for the model's inability to eclipse performance of the bigram model alone, and followed with results suggesting that the full model does better at appropriating higher likelihoods to test points that the language model component indicates are already somewhat likely.

Further, with respect to absolute metrics, the full model (and by extension, the bigram model) are highly accurate in their predictions. Across all test points, the median probability ranking of the correct behavior in the posterior predictive distribution for the best full model was third, and the correct behavior was in the top ten (out of 87) behaviors in 76.9% of the test events. Combined, all of these indicators give us confidence that our model is able to provide parameter estimates for EPA profiles that represent actual processes inherent in the data.

2.7.2 Perceptions during the Arab Spring

Behaviors

We now turn to a cautious interpretation of model results. Our focus is on parameters that give insights into how the English speaking news media portrayed the entities in our dataset during the Arab Spring. All results in this section are from parameter estimates of the full model run with three latent senses on the entire dataset. While the model performed slightly better with five latent senses we chose to use the model with three latent senses for parsimony. Qualitative conclusions are similar for both models.

Figure 2.4 displays 95% confidence intervals, as determined using μ_0 and σ_0^2, for the EPA profiles for all behaviors not already in the existing ACT dictionaries. Behaviors are ordered from left to right by their mean evaluative score, as this dimension tends to be the easiest to

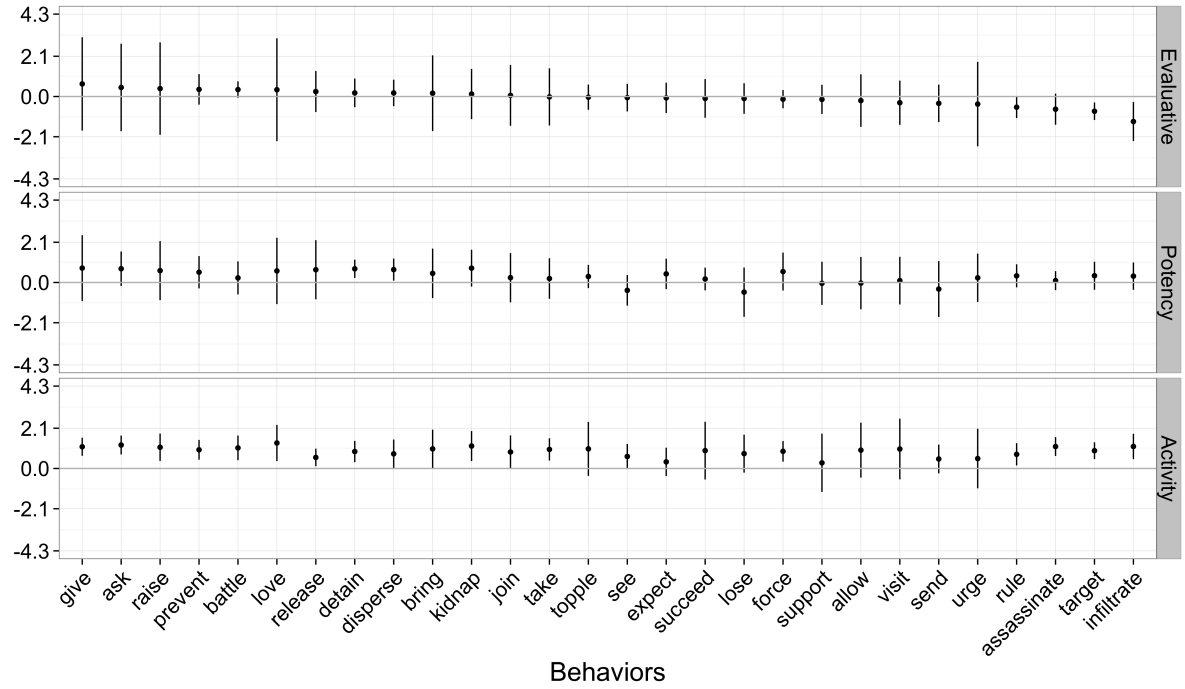


Figure 2.4: EPA Profiles for all behaviors used by the model *that were not already in the ACT dictionaries*. Confidence intervals are 95% intervals based on σ_0^2. A horizontal grey line is drawn at y=0 to delineate positive from negative values

conceptualize. Importantly, results are shown only for latent senses having more than 10 samples, as including data from latent senses with fewer than this number made the plot difficult to read and also displayed data that was heavily influenced by random initial values of m_0 and s_0.

Figure 2.4 shows that the model inferred a single, dominant latent sense for all behaviors - only a single latent sense had more than 10 samples for each of the behaviors listed. Thus, the model believed that across all cultural domains incorporated in the news data, the behaviors of interest had a relatively stable meaning. There are multiple reasons why this could be the case. From a mathematical perspective, GMMs operate to a certain extent on a "rich-get-richer" phenomena in that large clusters tend to attract more points. This may have an impact on the extent to which all behaviors clustered along one latent sense. However, as there are many identities that the model estimated to have multiple latent senses, it is plausible that other reasons exist for this observation. One possible reason is that behaviors simply have relatively stable and universal meanings across cultures. This finding may help to ground ACT analyses across cultures in the future.

Given the stability of the affective meanings of behaviors, we would expect that the EPA profiles of these terms fit our intuitive sentiments. We observe this to generally be the case, particularly for extreme values. The most positive behaviors in terms of mean evaluative score, "give", "ask" and "raise", intuitively seem to be things that a good identity would engage in towards another good identity. In contrast, "infiltrating", "targeting" and "assassinating" are indicative of behaviors that have a bad connotation. Similarly, behaviors at the higher end of

the potency spectrum, "give", "ask" and "kidnap", are behaviors that more powerful identities could engage in towards lesser individuals, and the least powerful behaviors, "lose" and "see", are relatively powerless. Finally, while all behaviors reported by the media and captured by the model are, unsurprisingly, reasonably active, "love", "ask" and "kidnap" can be considered to be three of the more active ones.

Though any analysis of such results is almost by definition subjective, the model's views on behaviors at the ends of the EPA spectrums fit with at least our own intuitions. When considering the "middle" of these distributions, however, two findings are also of interest. First, most behaviors are given more neutral values than what we expected. In comparison to data in the ACT dictionary used in the present work, the values are indeed slightly more neutral. For example, the 95% bootstrapped confidence interval around the Evaluative dimensions of the behaviors in Figure 2.4 is [-0.21,0.08] (mean -.05), a distribution which puts slightly more of its probability weight near zero than a similar band placed around all behaviors in the ACT dictionary [-0.01, 0.25], (mean .12). The cause of this is not immediately clear, and we return to this in the following section.

Second, we observe that there do exist behaviors for which results do not fit our intuitions. For example, kidnapping is actually a slightly *positive* behavior. While we cannot rule out other factors, the explanation for this seems to reside in what was considered newsworthy behavior during the Arab Spring. Although there are several instances of bad identities kidnapping good identities (e.g. "gunman kidnap woman"), the majority of the social events that involve kidnapping in our dataset are ones in which a good (or ambiguous) identity kidnaps another good identity ("father kidnap mother", "police kidnap child"). These events are newsworthy precisely because they are *un*expected (we would not generally expect fathers to kidnap mothers). Given information that good identities kidnap other good identities, however, the model is led to believe that the dominant sense of kidnapping is one of slightly positive evaluative sense.

This observation does not detract from the utility of the proposed approach - although this meaning for kidnap is unexpected, it is supported by this dataset. Future work will need to consider how to remedy, theoretically or methodologically, these differences between what is newsworthy and what is not. For example, one methodological remedy would be to modify assumptions about deflection. As newspaper articles likely include both culturally consistent information and more surprising, high-deflection events events, a bimodal distribution for deflection may be a more appropriate.

The fact that the model can estimate deviations around mean values for EPA profiles helps us to understand the extent to which the model is certain of its estimates. In the case of kidnapping, and many other behaviors in Figure 2.4, we see that the model is relatively uncertain of, for example, the "goodness" or "badness" of the behavior. Thus, for ambiguous cases (like kidnapping), the model responds with large deviations. The increase in this deviation may unfortunately lead to the masking of the existence multiple latent senses in the data, as these two sources of variation are in direct conflict during model inference. Future work will consider how best to account for this. However, current inference of both at least allows us to better understand how certain we can be of the different parameter estimates while still retaining the necessary theoretical components of ACT.

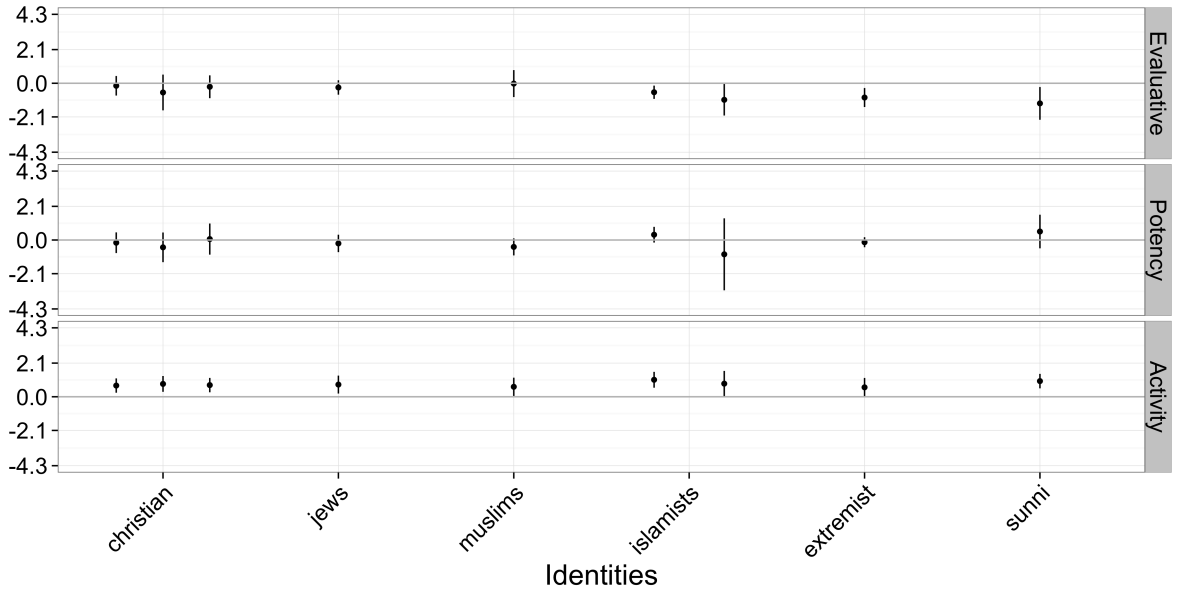


Figure 2.5: EPA Profiles for six identities of interest. Confidence intervals are 95% intervals based on σ_0^2. A horizontal grey line is drawn at y=0 to delineate positive from negative values

Identities

On some level, our analysis of the behaviors used by the news media was another exercise in model validation, as we observed that parameter estimates simply matched, for the most part, our intuitions. We now turn to an analysis of a small portion of the identities of interest to us and how they were perceived by major English speaking news media outlets. As a detailed analysis of all 102 identities in our dataset is not feasible, we choose to focus on one of particular interest, specifically those relating to religious groups[11]. Figure 2.5 displays results for these six identities in the same fashion as Figure 2.4. Note that several of the identities were used in multiple latent senses more than ten times. Thus, certain identities have more than one EPA profile associated with them.

The identities portrayed in Figure 2.5 represent the three major religious groups in the Middle East (Judaism, Christianity and Islam) as well as the identities "Islamist", "Sunni" and "Extremist". Unfortunately, we did not have enough data to compare perceptions of Sunni Muslims to Shiite Muslims. We hope to determine some way of doing so in the future. The figure shows that the media outlets found in our dataset collectively had a relatively neutral perception of the evaluative nature of the generic Muslim identity. Muslims were considered to be almost exactly neutral, with a mean evaluative score of -.02. Being neutral and reliably powerless [-0.96, 0.10], the Muslim identity in general was thus portrayed as more the victim than the perpetrator by the English speaking news media. This identity was even more neutral than Jewish and Christian identities. In fact, Figure 2.5 shows that Jews and Christians were frequently viewed as being slightly "bad".

[11]Results for all identities can be found online, at `https://www.dropbox.com/s/oas8rvlzgw4o6dj/all_identities.png?dl=0`

This result alone suggests that the news media did not focus on the religious aspects of the Arab Spring at the level of global religious identities. However, Figure 2.5 suggests that, far from identifying all Muslims as a neutral identity, the news media instead used more specific Muslim identities to connote a strongly negative view of the Muslim identity during the Arab Spring. The identities of Islamists, Muslims who believe politics should be driven by Islamic principles (Cammett and Luong, 2014), and Sunnis, the majority sect of Islam in the Arab world, are almost entirely negative and active. In all but one sense in which the Islamist identity is portrayed, these identities are viewed as being powerful as well. Even compared to the generic "extremist" identity, the Sunni and Islamist identities were strongly vilified.

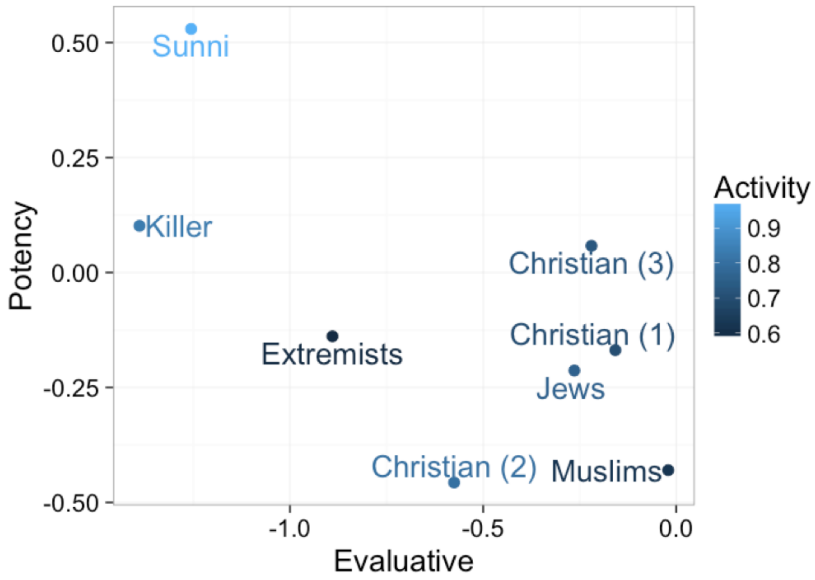


Figure 2.6: EPA Profiles for six identities. On the yaxis is the potency dimension, the xaxis is the evaluative dimension, color represents activity. Where there are numbers next to an identity label, that is a case where the model inferred multiple different dominant stereotypes, if theres no number the model decided that there was only one dominant stereotype.

It is relatively well-known that Islamist actors took advantage of the Arab Spring revolutions to gain power (Bradley, 2012), and that their ideological stance on government conflict with Western ideals of the separation of church and state. Figure 2.6 shows an alternative, three-dimensional (with color being one dimension) view of a select set of identities that provides another view of the Sunni identity. This plot makes it easier to see that Sunnis were represented as being quite bad, as bad as killers and worse than extremists, very active, and powerful. The data suggests that news agencies often portrayed Sunnis as the bullies of the Arab Spring.

Events on the ground certainly justify this to some extent- Sunni governments and militants acted oppressively in many situations. Also, certainly, the model or how we extracted data could obviously have its own biases. But it still seems possible as well that Sunnis as a whole were almost over-villified by the news media, perhaps due to implicit biases about Muslims that manifested in this more specific, culpable Muslim identity. Certainly we cant answer these questions based on the data or methods here, but this finding is an interesting point for future

work.

2.8 Conclusion

In the present work, we introduced a new methodology that can extract social events from text and then use these social events to extract sentiment profiles for identities and behaviors. The chief contribution is a statistical model that, by using the concepts of Affect Control Theory (Heise, 2007), provides insight into the soft, affective constraints that influence how we perceive and enact social events. This represents one of the first attempts to ground the analysis of sentiment in text towards generalized identities and behaviors in a way that uses the rigorous theoretical ideals put forth by ACT. From an NLP standpoint, our approach is one of the first efforts to extract a multi-dimensional sentiment profile for concepts; moving beyond the traditional approach of evaluation along a single, evaluative dimension. Further, our work allows for the extraction of multiple sentiment profiles for the same concept within a single corpora.

After describing our model, we provided a case study of its use on how the news media perceived and portrayed identities and behaviors of interest during the Arab Spring. Two findings were of interest. First, while the model found several cross-cultural differences in sentiments of identities, sentimental meanings of behaviors were universal across data from a large number of English-speaking news outlets across the world. While more work is need to better understand this finding, the possibility of stable meanings of behaviors across cultures would be of significant use in anchoring studies of cross-cultural and inter-group identity meanings.

The second finding of interest from our case study was that more specific, connotative Muslim identities of Sunni and Islamists were vilified by major English speaking news outlets, whereas the generic Muslim identity was considered to be rather neutral, even in comparison to the Jewish and Christian identities. A complete understanding of these perceptions requires a detailed consideration of both the events that actually occurred on the ground as well as an understanding of how particular events were perceived by the news media. In addition to those mentioned above, other well-known factors can be expected to have played a role in this finding. These including but are not limited the perpetual Sunni-Shiite conflict, the majority position of Sunnis in the Arab world and their resulting role in the revolutions. However, a comprehensive analysis of the relative influence of each of these factors, and how the Western media was influenced by them, is beyond the scope of the present work.

In taking a step beyond present methodological boundaries in a variety of fields, we made a host of decisions that had implications on our results. This was particularly true of our approach to event extraction. Three limitations should be noted in our current approach to extracting social events from text. First, we do not currently use the full subject or object, deciding to only use the single dependent term of a possibly multi-term entity (e.g. we only use "America", where the full subject might be "United States of America"). Similarly, we also ignore both settings and modifiers, and thus may lose significant semantic meaning. We assume these errors to be random at this point, and therefore that they "wash away" during model estimation. Second, in ignoring social events which do not have any existing ACT terms, we are removing a large set of potentially useful social events from our data. Future work is needed to better extract social events. Finally, we ignore the order of events in a document and over time. Accounting for temporal sequences might allow for a more accurate predictive model that does not rely on the

assumption that all social events extracted begin with the same transient meanings.

Aside from social event extraction, limitations also exist in our statistical model, including the heuristic way in which model parameters are initialized and its relatively poor performance on unlikely events. As we observed in our results, the model also seems to be slightly biased towards neutral sentiments, something we are currently working to understand. Combined, these limitations suggest that, as used here, the current iteration of the model is useful as a descriptive tool, providing insight from large amounts of data that can then be used for more focused, specialized studies. This use of the model is therefore similar to how current NLP tools, such as LDA (Blei et al., 2003), are used in the sociological literature, and we hope that our model provides another methodology for interpreting information from widely available textual data sources.

In the future, we intend to improve model performance in several ways. First, we can use slice sampling on hyperparameters (Neal, 2003) and assume a Dirichlet Process (Teh, 2010) on the number of latent senses per entity to better formalize the notion that the number of latent senses for each entity is unknown and thus should be estimated from the data. Similarly, we could extend the mixture model to use other covariates in the data (such as the particular newspaper from which a social event was extracted), similar to the recent efforts of Roberts et al. (2013). This work will aid in understanding the origin of different perceptions. As our model is agnostic to the source of social event extraction, we also hope to extend our efforts to consider different media beyond newspaper data.

The model could also be improved by a stronger relationship between the language model, which extracts semantic relationships, and the mixture model, which extracts affective relationships. We are currently exploring how the relationships between semantic "constraints" (Heise and MacKinnon, 2010) and affective "signals" can be formalized via cognitive modeling, an approach which may provide novelty in both the theoretical and methodological domains. Finally, we note that the ability to extract affect and relate it to behaviors is also an important extension to network text analysis (Carley, 1994). Future work could extend our model to this domain, perhaps utilizing ideas from relational event modeling (Butts, 2008), to extract more meaningful, valanced links between actors and between actors and objects and thereby expand the use of news data for network analytics.

The present work tackles a methodological question at the intersection of a variety of domains. Perhaps most importantly, we extend methodology surrounding Affect Control Theory to allow for the automated extraction of EPA profiles of entities from text. Given the vast expense of surveys in obtaining this information, the efforts described here, as well as those that build off them in the future, will lead to a stronger and more cost-efficient means of understanding how individuals perceive the actions and identities of others and how such affective constraints affect the way we think about and act towards others. Our work also provides researchers with the opportunity to perform historical analyses where, of course, surveys are not available. To this end, we intend to continue the public development of both code and documentation in a way that allows others to extend our work and use it without a strong programming background. The version of the code used for the present work is available at `https://github.com/kennyjoseph/act_paper_public`.

Chapter 3: Extracting Identities from Text

3.1 Introduction

Social scientists have long been interested in how and when identities are applied, used and created. Given the denotive and affective meaning identities convey, the specific one we choose to describe a person has a significant impact on the way others will act towards her (Heise, 2007). For example, because liars are "bad", we are unlikely to seek out a friendship with someone that we know has been labeled a liar by others. As social beings, we are implicitly aware of the importance of our identity, and are, consciously or not, consistently managing it in order to appear "worthy" of desirable identities in particular contexts (Goffman, 1959).

The study of identity is also prevalent in the natural language processing (NLP) community, though it tends to go by different names. Fine-Grained Named Entity Recognition (FG-NER) (Del Corro et al., 2015; Ekbal et al., 2010; Fleischman and Hovy, 2002; Hovy et al., 2011; Lin et al., 2012; Ritter et al., 2011; Yao et al., 2013) is focused in part on the problem of determining from a large set of identities which are most appropriate for specific individuals. Concept-level sentiment mining techniques have been applied to identity labels to understand the affective meaning they carry (Ahothali and Hoey, 2015; Joseph et al., 2016a). Finally, the authors of (Bamman et al., 2014) extract general semantic characteristics of particular "personas", which are similar to identities, from movie reviews.

Surprisingly, however, there does not seem to exist work that considers an even more basic question than those posed above- how do we capture the set of all identities that exist in a particular corpora? While a variety of heuristics have been applied to bootstrap lists of identities for FG-NER (e.g., (Yao et al., 2013)), to the best of our knowledge, the following prediction problem has not yet been explored:

> *Given a set of text data, label each word in the text as being representative of a (possibly multi-word) identity*

From a sociological perspective, even a method for extracting where identities are used in a given corpora would provide a new way to study their semantic and affective properties. For example, MacKinnon and Heise (Heise, 2010a), two prominent social psychologists, argue in their recent book on identity that understanding the structuring of identities into semantic clusters using text analysis can help us to understand how individuals navigate social life and how cultures as a whole create taxonomies of identities (e.g., into those related to occupations versus those related to family).

The first contribution of the present work is a supervised classifier that addresses the NLP task posed above for tweets. We first sample 1000 tweets from a large, domain specific corpora and annotate it with labels indicating which terms represent identities. As we are the first

to approach this problem, we use an iterative coding scheme and consult with identity scholars to derive theoretically grounded rules for the labeling process. We then construct features derived from both standard lexical structures and from a variety of tools recently produced by the NLP community around Twitter. In particular, features are derived from the output of a Twitter-specific dependency parser (Kong et al., 2014) as well as from word-vectors trained on a large Twitter corpus using the GloVe algorithm (Pennington et al., 2014). Additionally, we make use of existing dictionaries of identity labels and also construct a bootstrapped dictionary of identities from unlabeled data via the use of high-precision lexical patterns.

Model performance is first tested on this set of 1000 tweets using cross-validation. As an additional step to assess the quality of our predictions, we also obtain and label an additional 368 tweets made public by other NLP researchers and use them as a validation set. On this validation set, model performance is compared to a rule-based, dictionary-based approach. F1 scores exceed .75 for the full proposed model and outperform this baseline by 33%.

The second contribution of this work is a case study that demonstrates one particular opportunity for sociological research that our method allows. We apply our trained classifier to over 750M tweets sent by 250K Twitter users who were actively engaged in discussion of the recent deaths of Michael Brown and Eric Garner. Our case study explores the following question: *how do the identities in our dataset cluster into semantically organized sets of identities?*.

We provide both quantitative and qualitative analyses of identity sets, or clusters, that result from applying latent Dirichlet allocation (LDA) (Blei et al., 2003) to a user by identity matrix extracted from our corpus. Encouragingly, the resulting clusters line up well with both prior work and intuition. We find several clusters of identities that match those found by Heise and MacKinnon (Heise, 2010a) in their related work and also find clusters that align strongly to contemporary social issues (e.g., the Arab/Israeli conflict). We then briefly explore how these two "types" of identity sets can be differentiated based on their affective meanings. We also consider how differences in social contexts may compel individuals to utilize particular identity sets more often than others.

3.2 Literature Review

We divide our review of the literature into two parts, one focusing on the sociological literature on identity and the second on related research in the NLP community.

3.2.1 Sociological Literature

Smith-Lovin (Smith-Lovin, 2007) defines identities as the ways in which an individual can label another person with whom she has had an interaction. Smith-Lovin continues to define three general types of identities. Role identities indicate positions in a social structure (e.g., occupations). Category identities come from identification with a social category, which are "inclusive [social] structures that require merely that all members share some feature" (Cikara and Van Bavel, 2014) (e.g., race, gender). Finally, social identities indicate membership in social groups, a collections of individuals who a) perceive they are in the same social category, b) share a common understanding of what this category represents and c) attach an emotional meaning to this category.

The broad definition of identity provided by Smith-Lovin foreshadows various difficulties in developing a methodology to extract them from text. While we provide a more nuanced

discussion of practical difficulties in Section 3.4.1, the chief socio-theoretic issue relates to the difficulty of distinguishing between a "compound" identity, one that has multiple words, and a single identity that is modified. For example, the phrase "black woman" could be viewed as the identity "woman" modified by the adjective "black", or as a single identity, "black woman".

From a linguistic perspective, Recasens et al. (Recasens et al., 2011) suggest that identity should be considered to be both relative and to be varying in granularity, and thus a determination of the granularity of interest should absolve us from these problems. A complementary sociological perspective can be drawn from the theory of intersectionality, which emphasizes the importance of understanding social categories as being social constructions (Saperstein et al., 2013) and thus the importance of defining which identities are or are not compound via social consensus. While much remains to be done along these lines theoretically, our labeling scheme attempted to utilize a small number of agreed-upon intersectional identities from the literature. At the same time, where intersectionality was non-obvious, we adopted a coarse-grained labeling approach, looking only for root-level identities and leaving identification of modifiers to future work.

Regardless of the definition of identity used, MacKinnon and Heise (Heise and MacKinnon, 2010) perform the only attempt we are aware of to enumerate identity labels on any large scale. Their efforts come in two parts. First, they extract all terms from WordNet (Miller, 1995) that are lexical descendants (recursively) of the term "human being" and then perform a qualitative analysis to understand taxonomic structure in the resulting sets of identities. Their work suggests a set of twelve categories of identities that include, for example, occupation and religion. The authors then perform a semantic analysis of identities in an offline, professionally written dictionary, where they cluster a semantic network extracted from the dictionary to obtain a similar collection of identity sets. These structures are referred to as *institutions* - for example, one institution includes identities such as siblings and parents, representing the institution of family and marriage (pg. 79). Institutions thus consist largely of semantically coherent sets of identities that are applied together in specific social contexts.

The approaches taken by MacKinnon and Heise (Heise and MacKinnon, 2010) to define identities on a large scale and to uncover semantically coherent sets of these identities has certain advantages. In particular, both WordNet and the professional dictionary are human curated and widely used, suggesting a high level of precision in both semantic relationships and identity labels used. However, the approach also has disadvantages. First, the datasets used are curated by a specific collection of individuals whose views may not be entirely reminiscent of social consensuses on identities or their meanings. Second, the datasets that they use base semantic relationships largely on denotive meanings of identities. As we will see, affective meanings can be equally important in our understanding of semantically coherent clusters of identities.

3.2.2 NLP Literature

The task of *Named Entity Recognition* (NER) is defined by the goal of extracting entities from text and categorizing them into a general typology, most often into the categories of People, Locations and Organizations. One of the earliest applications of NER to Twitter data is the work of Ritter et al. (Ritter et al., 2011), who develop and test a semi-supervised model based on Labeled LDA (Ramage et al., 2009). Ritter et al.'s work moves beyond the simple Person, Organization, Location classification to finer-grained classifications of entities, and is thus one

of, if not the, earliest application of FG-NER to Twitter.

Research on FG-NER uses large sets of entity labels and tries to apply them to, for example, people. As opposed to labeling "Michael Jackson" as a Person entity, an FG-NER system might label him as a "musician". Entity labels used to classify people are by definition identities. It is thus unsurprising that recent work in FG-NER (Del Corro et al., 2015; Ekbal et al., 2010) uses WordNet to construct lists of entity types in a similar fashion to the work of MacKinnon and Heise. Because of this connection between entity labels and identities, features used in FG-NER models are applicable to the present work. Of particular interest are the feature sets utilized by Hovy et al. (Hovy et al., 2011) and del Corro et al. (Del Corro et al., 2015), which are derived from lexical patterns (e.g., typing Steve Young with the label quarterback given the text "Quarterback Steve Young"), dependency-parses, parts-of-speech and word-vectors, all of which are similarly utilized here.

Yao et al.'s (Yao et al., 2013) recent work in the FG-NER domain is perhaps most relevant to the work here. The authors develop an approach to type entities with labels from free text. In their work, the "type system", which is loosely equivalent to the set of identity labels we wish to construct, is generated from a pattern-based extraction method from text. After constructing this dictionary of entity types, a matrix factorization method is developed to apply these types to Named Entities.

While our work is thus in many ways related to FG-NER research, it is important to observe that the problem we are interested in has a fundamentally different goal. Whereas in FG-NER, one is attempting to find appropriate labels for Named Entities, here we attempt to find all labels that are used to describe any human or set of humans. This distinction is important for two reasons. First, NER systems assume that entities can be labeled with factual types. However, in highly emotional situations, like the case study considered here, it seems unlikely that such factual labels will be prevalent or even interesting, particularly in Twitter data. Rather, as Bamman has noted (Bamman, 2015), what is interesting is how different identities are found in text based on the current social context of the individual writing the text. Second, of particular interest to us is how identity labels themselves are related to each other, not how they apply to entities. While it may be interesting to understand how, for example, different labels are applied to Michael Brown, the current work is focused largely on how people view generic groups of others (e.g., the police) that are rarely regarded as entities.

Our case study is similar to a variety of recent efforts to perform topic-modeling on Twitter (Yang et al., 2014; Zhao et al., 2011). Additionally, our focus on affective meaning is relevant to more recent approaches that combine sentiment analysis with semantic connections between terms (Hoang et al., 2014). The efforts in the present work complement this line of research by considering a particular kind of topics- specifically, "topics"[1] of identities. While the present work uses straightforward methods to perform this clustering, we look forward to leveraging more complex models that account for, e.g., spatio-temporal properties of the data in the near future (Ahmed et al., 2013; Eisenstein et al., 2010; Wei et al., 2015b).

[1] or preferably here, "clusters" or "sets"

3.3 Data

A variety of data sources were used in the present work. Here, we give a brief description of each. All code used to collect data, all dictionaries mentioned and all labels for the supervised problem, as well as all code to run the models to reproduce results, will be made available at `http://github.com/kennyjoseph/identity_extraction_pub`.

3.3.1 Twitter Corpus

On August 9th of 2014, Michael Brown, an unarmed 18-year old African American male, was shot to death by Darren Wilson, a member of the Ferguson, MI police department. Over the next few days, questions began to arise surrounding the circumstances of Brown's death. Over the next several months, two important series of events played out. First, a grand jury was organized to determine whether or not to indite Officer Wilson for any charges related to the death of Michael Brown. Second, a host of mostly peaceful protests were carried out on the streets of Ferguson and elsewhere, demanding justice for yet another young black male that they believed had been wrongly killed at the hands of a police officer.

On November 24th, the grand jury determined there was no probable cause to indite Darren Wilson for any crimes related to the death of Michael Brown. This decision was met harshly by critics both online and on the streets of cities around the United States. Less than two weeks later, another grand jury, this time in Staten Island, also chose not to indite a white police officer over the death of Eric Garner, another black male. Garner's death, which was notably caught on video, reignited flames from the protests in Ferguson, both online and in the streets and from those that both condemned and, unfortunately, those that celebrated the deaths of Garner and Brown.

The tweets used for the present work are a subset of a corpus of approximately two billion tweets from around one million Twitter users who we considered to have been an active participant in these discussions. From August, 2014 through December, 2014, we monitored the Twitter Streaming API with a variety of keywords that were relevant to events in Ferguson following the death of Michael Brown and events in New York City leading up to and following the trial resulting from the death of Eric Garner. An active participant is defined as any user who sent more than five tweets that were captured in this sample. For these active users, we collected their full tweet stream[2]. For the present work, we focus on a subset of users that we expect to be both human and to be active on the site. Specifically, we focus on users who have sent between 50 and 15K total tweets, have less than 25K followers and that have been on the site for 2 or more years. From this set, we consider only English language tweets[3] without URLs[4] that have five or more tokens.

For the purposes of developing our classifier, we extracted 1000 non-retweets from this set of filtered tweets. We ensure that this sample contains at most one tweet per unique user. Because we expected tweets with identities to be relatively rare, we used three methods to over-sample tweets with identities in them. First, we use Vader (Hutto and Gilbert, 2014), a sentiment classifier, to extract only tweets with some form of sentiment. This is because affective relationships

[2]Up until their last 3200 tweets, as allowed by the API

[3]determined with the languid library (Lui and Baldwin, 2012)

[4]as determined by Twitter

between identities tend to have an affective component (Heise, 2007)[5]. Second, we ensure that 10% (100) of the tweets we sampled had one of twenty generic identities labels (e.g., bully, husband) drawn from one of the identity dictionaries described below. Finally, because we were specifically interested in views on the police, we ensure that 15% (150) of the tweets had the word "police" in them.

Due to the large extent to which we utilized sub-sampling, and the fact that our corpus selects on a very distinct dependent variable (Tufekci, 2014), it was necessary to obtain an outside dataset to validate the model. We chose to use the corpus discussed in (Owoputi et al., 2013). Of the 547 tweets in the corpus, only 368 of them were still able to be extracted from the API (i.e., the rest had been deleted or sent by a user who had since closed their account). We hand labeled each of these 368 tweets and use this set as a validation set.

3.3.2 Dictionaries

A variety of dictionaries, or word lists, of identities already exist. We leverage several of these dictionaries here. First, Affect Control Theorists, who focus on culturally-shared affective meanings of identities and their behaviors, maintain an open-source listing of identities used in their survey studies (Heise, 2010a). As noted above, WordNet contains an implicit identity dictionary, which can be constructed by collecting all terms that derive from the "Person" term. In addition to these two lists of general identity terms, we also adopt dictionaries for specific types of identities we expect to be prevalent in our dataset. From the GATE (Fleischman and Hovy, 2002) set of gazeteers, we utilize a listing of occupations (frequently used as role identities) and nationality-based identities. Finally, we obtain a set of racial slurs for a variety of races from the Racial Slur Database[6]. In total, we thus have five distinct lists of identities.

In addition to dictionaries for identities, we also utilized dictionaries for non-identity words. We drew these from the GATE set of gazeteers, as well as from the set of dictionaries from the Twitter NLP library[7], although several dictionaries from each were excluded based on manual inspection. We also use all terms in WordNet that do not derive from the Person entity as a non-identity dictionary. Finally, we include a generic stopword list.

3.3.3 Word Vectors

As opposed to allowing the model to learn from one-hot encodings of unigrams, we opt for the dense representations of words afforded by recent work in representation learning (Mikolov et al., 2013b). Specifically, we leverage a large set of 50-dimensional word vectors that have been trained on a large set of Twitter data using the GloVe algorithm[8] (Pennington et al., 2014). As opposed to a unigram-based approach, these dense word vectors allow the model to learn general classes of words that are identities rather than specific unigrams.

[5]While there is a danger that this may bias our classifier towards extracting identities only from sentiment-laden tweets, there was nothing in model output that lead us to believe this to be the case

[6]`http://www.rsdb.org/`

[7]`https://github.com/aritter/twitter_nlp`

[8]available for download at `http://nlp.stanford.edu/projects/glove/`

3.4 Methods

3.4.1 Labeling Process

For the present work, each term in each tweet was labeled as being either inside or outside of an identity label - an "IO" labeling scheme. Labeling of the data was completed in two steps. In the first, a set of thirteen annotators, most unfamiliar with the project, were each asked to annotate around 150 tweets, giving us two labels for all 1000 tweets. Annotation was performed using the brat rapid annotation system (Stenetorp et al., 2012). Guidelines gave annotators an expanded definition of identity as compared to the one presented in this article, as well as a variety of examples. We also asked annotators not to label individuals or pronouns as identities (including modified forms, e.g., "people I know"), not to include organizations themselves as identities (e.g., "Congress announced today" would not have any identities, while "A member of Congress" is an identity), and to only label full words as identities (e.g., "#blacklivesmatter" would *not* be labeled as an identity).

While the guidelines given left us with a very limited definition of what constituted an identity, such a limited definition was necessary for the task to be completed with any amount of agreement. As identity labels were sparse, we only chose to evaluate inter-annotator agreement on tokens where at least one annotator claimed that the span of words was a part of an identity label. On such tokens, agreement was 67%, that is, if one annotator labeled a particular span of tokens as being an identity, there was a 67% chance that the annotation matched exactly with the second annotator.

In reviewing these annotations, we found three main sources of disagreement. First, a few annotators, although told that they were being ignored, still labeled pronouns as identities. Second, annotators varied in the extent to which they included modifier terms in their annotations. This was particularly the case where identities served to modify individual persons (e.g., in the phrase "Mayor de Blasio", "mayor" is a modifier and thus not of interest as an identity in the present work). Finally, identity words that were used as generic, emphatic statements in the text ("man" in "Come on, man!") were differentially labeled by annotators. We eventually chose to ignore these, as they serve largely as general pronouns rather than statements about identity.

After resolving these general themes of disagreements between annotators and fixing annotation errors, we consulted with identity theorists to finalize any additional rules and guidelines for annotations. Once firmly established, we reviewed all annotations to confirm their adherence to these guidelines and also applied them to the additional 368 tweets from outside the corpus.

3.4.2 Creation of Bootstrapped Dictionary

Existing identity dictionaries did not include many of the identity terms in our labeled data- in fact, even though they contained over 11K entries, they captured only 64% of the identities in our labeled tweets. While, of course, a statistical model allows us to generalize beyond these dictionary terms, another important source of information to leverage was the set of unlabeled tweets in our corpora. A common usage of unlabeled data in FG-NER studies is to extract possible entity types for individuals by *bootstrapping* a dictionary using high-precision lexical patterns (Del Corro et al., 2015; Yao et al., 2013). Generally, these bootstrapped dictionaries are created by first performing coarse-grained NER on the data, and then using lexical patterns like

"I Am" Rule	"Person" Rule	Not in Dicts
girl	black person	mess
man	wrong person	human
bitch	young person	legend
kid	favorite person	joke
guy	old person	pussy
idiot	nice person	thot
asshole	beautiful person	blessing
woman	amazing person	nightmare
boy	bad person	disgrace
h*e	real person	cutie
friend	innocent person	texter
baby	stupid person	goddess
keeper	homeless person	g
$$$ga	random person	old

Table 3.1: Three lists of terms from the bootstrapped dictionary, sorted by frequency of occurrence. On the left, top terms from the "I am a", "he is a" etc. ruleset. In the middle, top terms from the "[Identity_Label] person" rule. The final column gives the 15 most frequent phrases extracted from the "I am" ruleset that were not in any existing identity dictionary we used

"[Entity_Type] such as [Entity]" to extract entity types.

Unfortunately, most of the prior work is focused on news or web data, and many of the patterns that were used in these sources of data, such as appositional phrases ("Joe, the dentist"), almost never occurred in our Twitter corpora. Further, NER on Twitter data is still a very difficult and time-intensive problem (Ritter et al., 2011). Consequently, we adopted a pair of slightly modified lexical patterns from the existing literature to build our bootstrapped dictionary. First, as opposed to using NER as a precursor for label extraction, we instead use pronouns as the base of our patterns. We thus extract sets of tokens starting with "he is", "she is", "I am" or "you are" that were followed by the word "a" or "an" and consider the first noun that follows to be an identity (e.g., "liar" is extracted from "he is a liar"). Second, we found that in almost all cases, terms proceeding the words "person" or "people" (and variants of these words) were identities (e.g., "annoying person"). We thus extract all terms of the form "[X] people" or "[X] person" and add these to our bootstrapped dictionary as well.

We used our unlabeled corpus to extract all phrases matching these two sets of patterns, and kept a count of the number of times we capture each unique phrase using each pattern. After obtaining these counts from our full, 2B tweet dataset, we remove all phrases that occurred fewer than 10 times within one of these patterns. In total, we were left with 30.5K unique identity terms. Table 3.1 displays three columns that help to describe the resulting dictionary. The first column shows the fifteen most frequently captured terms from the "I am", "he is", etc. ruleset[9]. The second column shows the fifteen most frequent terms collected using the "person" ruleset.

[9]Note that $$$ will uniquely stand for the letters "nig" throughout the article due to the relevance of this set of identities. All other words we do not wish to print will be edited with *s

The final column shows the fifteen most frequent terms from the "I am" rule set that were not in any of the obtained dictionaries. As is clear, true identities (e.g., "cutie", "g") are mixed with noise words, like "blessing".

However, the resulting dictionary is nonetheless useful. In the 1000 tweets used for development and testing, 91% of all terms labeled as identities are found in this bootstrapped dictionary, and 96% are captured when we combine the bootstrapped dictionary with the existing dictionaries. This high level of coverage from our dictionaries allows us to focus heavily on precision.

3.4.3 Model Description

The prediction problem we address - determining whether or not each word in our text is an identity label or part of an identity label - is highly imbalanced. Only around 4% of the words in our labeled data are identities. Consequently, our modeling approach and our evaluation are geared towards techniques for imbalanced prediction problems. One such technique is to run a filter through the data to remove uninteresting words and thus reduce the imbalance. Consequently, we develop a two-stage model to predict, for each token in each labelled tweet, whether or not it was (or was part of) an identity label. The first stage of the model is a rule-based classifier that labels all stopwords and words not in any of our identity dictionaries as negative (i.e., as not containing an identity). Terms missed from our dictionary were generally stopwords themselves within multi-word, intersectional identities (e.g. "Of" and "The" in the identity "Leader Of The Free World", a direct reference to the American presidency). This filtering step, we believed, would help the statistical model to "focus" more on interesting and meaningful differences between possibly relevant terms, rather than learning to simply differentiate stopwords from non-stopwords. In practice, it had a relatively small effect on performance relative to our feature set.

The second step of the model applies a straightforward, L1-regularized logistic regression model on each term in each tweet individually. While we expected that a sequential model (e.g., a CRF) would perform better on the task at hand, we found that a per-term regression approach with a strong set of features tended to perform as well or better than the sequential models we tested. We expect that this may occur due to the fact that the majority of identities (85%) were only one word long.

The features for each word, W_i, we used in our model are given in Table 3.2. The table displays three columns. The first column provides feature names. The second column gives the words for which the features are created. For example, for each word W_i, we use the Penn Treebank POS tag for the word W_i itself, the previous word W_{i-1} and the following word, W_{i+1}. The final column provides additional information, if necessary.

Features fall into one of three general categories. First, we use standard lexical features. These features include coarse-grained part-of-speech (POS) tags using the tagset described in (Owoputi et al., 2013), as well as finer-grained, Penn Treebank style tags. Lexical features also include various traditional word form properties (e.g., is the word capitalized?) and the Brown clusters utilized by (Owoputi et al., 2013) in their POS tagging model. The second set of features includes vector representations of the word W_i itself, its head word in the dependency parse and the word that exists at the end of W_i's "chunk".[10] Importantly, the dependency parser

[10]Additional features signifying words that were outside of the vocabulary for the word vector data from (Pen-

Lexical Features		
Penn Treebank POS tags	W_{i-1}, W_i, W_{i+1}	e.g., NNP, VBP
Coarse-grained POS tag	$W_{i-2}, W_{i-1}, W_i, W_{i+1}, W_{i+2}$	e.g., V, N
Prefix/Suffix, length 1 and length 3	W_i	From "liar", the set l, lia, iar, r
First letter Capitialized, All Capitalized	W_i	e.g., "ALL_CAP"
Has digit, Is a Single Digit, Has a dash	W_i	e.g., "SINGLE_DIGIT"
Brown Cluster using data from (Owoputi et al., 2013)	W	"One-hot" vector encoding
Word Vector Features		
50-dimensional word vector	W_i	All zeros if word not in vocabulary.
50-dimensional word vector	Head of W_i in dependency parse	All zeros if word not in vocabulary.
50-dimensional word vector	Last word in chunk for W_i	All zeros if word not in vocabulary.
Dictionary Features		
Is in any existing identity or non-identity dictionary	W_{i-1}, W_i, W_{i+1}	Feature for name of each dictionary the word or its bi- or trigram is found in; e.g., $in\_dict\_wordnet\_identities$
In bootstrapped dictionary at a particular cutoff	W_{i-1}, W_i, W_{i+1}	Cutoffs of 1000, 10000 and 100000 are used ; e.g., $in\_bootstrap\_dict\_1000$
Is in stopword list	W_{i-1}, W_i, W_{i+1}	Generic stopword list for Twitter

Table 3.2: Features used in our statistical model

provided by (Kong et al., 2014) for Twitter sacrifices the semantics of the Stanford Dependencies in order to provide reliable accuracy, only providing rough dependency connections between words. Further, chunks are quoted as they are determined heuristically, largely by connecting consecutive noun phrases together. More advanced chunking approaches (Ritter et al., 2011) were too time consuming for our full dataset.

Finally, we incorporate features that use the afore mentioned dictionaries. More specifically, if a word or any bigram or trigram the word is in is found in a particular existing dictionary, we add a feature to the model to indicate this. So, for example, if the word "liar" were to be found in both the Affect Control Theory list of identities and the WordNet list of identities, it would have both the binary features $in\_dict\_wordnet\_identities$ and $in\_dict\_ACT\_identities$. We use this approach because various dictionaries showed various levels of noise, and using features

nington et al., 2014) did not change performance and were excluded.

differentiated by dictionary name improved the model.

We take a similar approach in our utilization of the bootstrapped dictionary, assuming that the more frequently a word is captured by our patterns, the more likely it is to be an identity in any given tweet. We create various frequency "cutoffs" and use each as a feature. Consequently, if the word "liar" were to be extracted 10000 times by our lexical patterns, it would have both the binary features $in\_bootstrapped\_dict\_1000$ and $in\_bootstrapped\_dict\_10000$. This use of coarse-grained cutoffs worked better for this problem than using the actual frequency value itself (or any transformation of it we tried). Note that we also include dictionary features for the words before and after W_i, and that we consider all unigrams and bigrams a word is in when looking in the bootstrapped dictionary (identities in the bootstrapped dictionary are at most two words long).

3.4.4 Model Evaluation

We evaluate model performance in two ways. First, we use cross validation on the 1000 labeled tweets from our corpus, where we tune the regularization parameter for the logistic regression and analyze model performance with various feature sets. Note that five-fold cross validation is performed with an 85/15 train/test split instead of an 80/20 split because we ensure that only the 750 tweets selected at random from our corpora (and not the 250 selected via a keyword search) are used in the test set. Doing a full cross validation would only served to artificially inflate model performance.

We then analyze performance of the best model on the validation set. In both cases, the outcome metric of interest is the F1 score[11] on the positive (identity) class. We use F1 as opposed to a simple accuracy score due to the imbalance between the classes.

3.4.5 Baseline Model

In order to provide a useful comparison of model performance, we develop a simple but effective dictionary and rule-based classifier as a baseline. The classifier works in a similar fashion to the filter described above, with two differences. First, the baseline model is tested with various subsets of the identity dictionaries, as opposed to simply using all words in all identity dictionaries. Second, it uses POS tags, only labeling nouns as identities.

In sum, the baseline model classifies any noun in the list of identities it is given as an identity and labels all other terms as non-identities. We run this baseline model with all possible combinations of dictionaries (e.g., all dictionaries by themselves, all pairs of dictionaries, etc) to find the strongest baseline model to compare to, "optimizing" for F1 score on the 1000 tweets not in the validation set.

3.5 Results

3.5.1 Model Performance

Figure 3.1 shows model performance on the five-fold cross-validation task. On the vertical axis of Figure 3.1 are the different feature combinations we tested, on the horizontal axis, the mean F1 score for the model with the optimal regularization parameter for that feature set. Error bars

[11]The F1 score is the harmonic mean of precision and recall

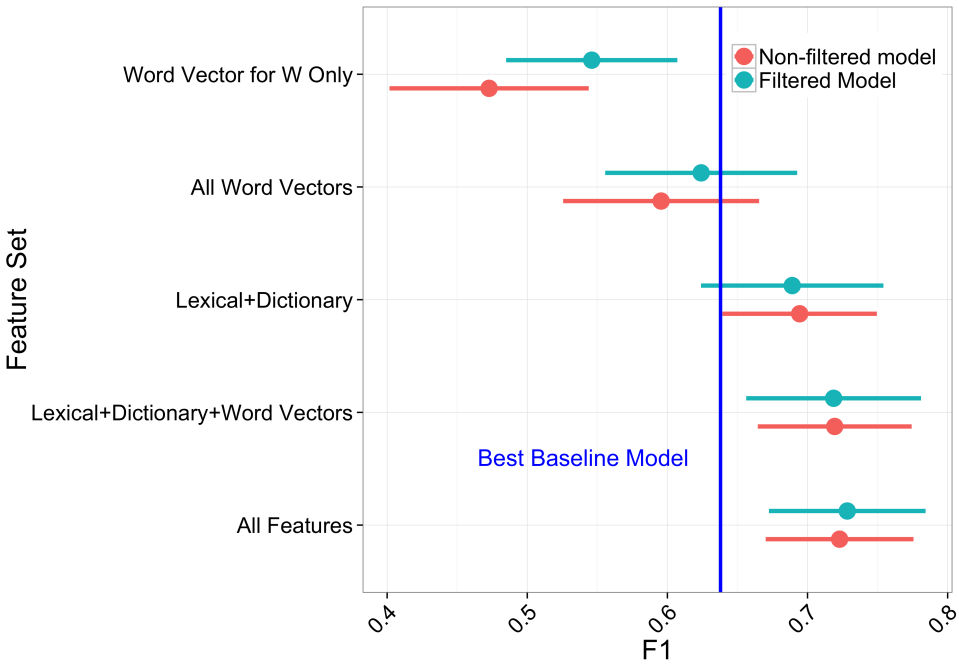


Figure 3.1: Cross-validation results. Different feature sets are given on the vertical axis, F1 score on the horizontal axis. Error bars are 1 standard deviation, different colors represent with/without the filtering step. The blue line represents the best dictionary, rule-based baseline

show one standard deviation, and results are given with and without the filtering step for each feature set combination. Figure 3.1 also displays a vertical blue line that depicts performance of the best rule-based baseline. As the rule-based model did not need to be trained, error bars are not shown- we simply ran it once on the entire dataset.

Figure 3.1 shows that lexical, dictionary-based and word vector features are all predictive on their own. Combining all of these features into a single model results in improved performance, though word vectors for the head term and final chunk word add only a small amount of additional information. The best performing model overall is the full model using the two-step filtering, with an average F1 score of .74. Taking the best full model and applying it to the validation set, we find it performs slightly better than during cross-validation, with an F1-score of .76, an improvement over the best baseline method (F1=.64) by 33%. This improvement over the dictionary-based methods on the validation set is almost double the improvement over the baseline on the Ferguson data (14% improvement, shown in Figure 3.1). This increase in the performance difference can be attributed to the fact that in running the dictionary baseline on the same data used to construct the dictionary, the baseline overfits to this particular dataset.

3.5.2 Error Analysis

Errors made by the model during both cross validation and on the validation set can be roughly categorized into four types. First, errors were made when terms typically used for identities were applied to non-human entities. For example the word “member” in the phrase “Big East member” refers to a university, not a person, that is a member of the Big East athletic conference.

Second, much like our original annotators, the model had difficulties distinguishing modifiers from identities, particularly when these modifiers were applied to people. For example, using our definition of identity, the term "quarterback" in "quarterback Steve Young" is *not* an identity. Third, the model occasionally mis-classified organization names as identity labels (e.g., "Citizen" in "Citizens United". Adding NER labels from the classifier developed in (Ritter et al., 2011) helped slightly, but substantially increased the amount of time it took to run the model. Finally, the model struggled with misspellings (e.g., "team-mate").

In sum, it can be said that the model's errors might affect the case study performed here in that it had a tendency to over-estimate the extent to which common identity words (e.g., "member" above) were actually used as identities and to under-estimate the number of times infrequent or misspelled words were used as identities. Given our fairly high-level focus below, we do not believe these biases to have had a strong impact on results.

The errors we observed signify two possible avenues of future work. First, continued discussions with identity theorists and linguists may help to better understand how to address certain labeling issues, in particular those related to modifiers/compound identities. It is likely that our labeling process still had inconsistencies, which detracted from our ability to learn patterns in the data. Second, our work will benefit from leveraging additional types of features, in particular ones that leverage verb-based patterns (e.g., (Grycner et al., 2015)).

3.6 Case Study - Ferguson Data

3.6.1 Overview

For our case study, we first trained our classifier on the full set of labeled tweets. We then ran it on tweets from 250K users in our dataset that fit the qualifications provided in Section 3.3.1 (User had 50-15K tweets and ¡25K followers, no retweets, tweets with URLs or non-English tweets). As a heuristic to identify compound identities, we combine all sequential identity terms in to a single identity phrase. Such labels were sufficiently sparse that we do not expect this decision to have a strong influence on the general results we present here. Of the 750M tweets sent by users in our case study, 6% (45.4M) both fit our requirements and contained at least one term our model identified as an identity. Visual inspection of the model's predictions on tweets from a handful of users suggests that precision and recall are roughly the same as in our validation set, around 75% for both precision and recall. Within the set of tweets that contained at least one identity, 94% of the terms were labeled as non-identity terms, an appropriate level of sparseness given what we observe in our hand-labelled data.

In total, our model extracts 145K unique identity labels from the text, around 14K of which occur more than ten times. This number is considerably higher than the number MacKinnon and Heise drawn from WordNet (5.5K), suggesting, unsurprisingly, that a significantly higher level of both social complexity and noise is fostered via our approach. Figure 3.2 shows the top ten identities discovered along with their frequency of use, as represented by the proportion of all identities that they account for. Nearly one in three times a user expressed an identity, it was one of the ten terms listed in Figure 3.2. The identities shown generally fit intuitions - they capture two of our most obvious physical traits, sex ("girl", "guy", "woman", "dude") and age ("child", "kid"), as well as our most important forms of relationship - friendship ("friend") and kinship ("mom", noting that "dad" is the 11th most popular identity). Further, as we know that sports are

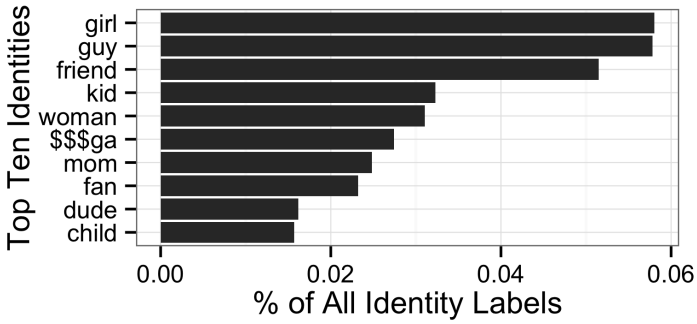


Figure 3.2: On the vertical axis, the top ten identities uncovered by the model. The horizontal axis shows the number of times this label was used as a percentage of the total number of identity labels in the corpus

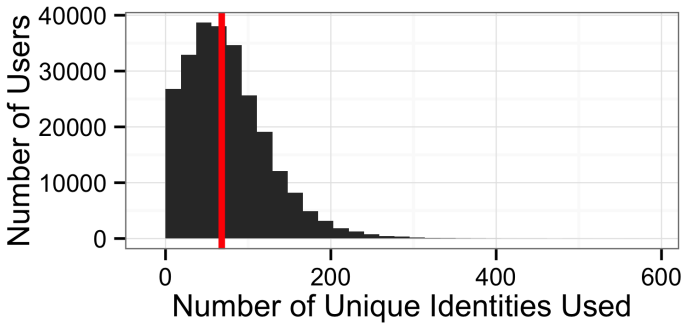


Figure 3.3: A histogram of the number of unique identity tweeted by each Twitter user. A red line has been drawn at the median of 68 unique identity labels

a popular topic of discussion on Twitter (Bosagh Zadeh et al., 2013), we were unsurprised to see frequent use of the identity "fan". Finally, given that we selected our sample based on a racially charged issue, we were also unsurprised that tweets in our dataset frequently contained the term "$$$ga".

Figure 3.3 plots a histogram of the number of unique identity labels expressed in tweets by each user in our dataset. The median user had *68 unique identity labels* in their tweets[12]. While in isolation, this number seems fairly high, users on average introduced a new identity label in only around 2.5% of their tweets. Table 3.3 displays the unique identities sent by one random user with exactly this number of identities - while again, 68 may seem like a large number, the identities displayed in the table seem quite reasonable. Further, on Twitter, where writing space is at a premium and audience size can be large, the succinct but widely-shared meanings that identities convey might lead to an even stronger reliance on them to share social information. Future work in other domains or with different types of models may therefore lead to a different estimate of this value. Regardless, it is clear that while people on Twitter tend to focus on only a handful of topical domains (Bosagh Zadeh et al., 2013), within each domain individuals perceive a rich typology of identities.

[12]A file containing all identities used, one line per user, can be found at goo.gl/ipcEuh

boyhoe	weirdo	adult	mom	nurse
freshman	cousin	poser	lady	teacher
kid	toddler	bestfriend	gramp	stranger
gf	professor	punk	brother	millionaire
youth	dick	chief	bastard	actor
guitarist	father	hairdresser	mason	virgin
woman	president	redhead	monarch	asshole
boyfriend	underclassman	boy scout	coach	dude
girlfriend	rapper	chap	bitch	female
president	girlfriend	hero	girl	prince
cop	manager	fan	dad	pilot
mother	guy	hater	friend	homosexual
whore	clown	tender	bunny	mcm
kate	person	guysss		

Table 3.3: Unique identities used by one random Twitter user with exactly the median number of unique identities over all users. Possible false positives are shown in red

3.6.2 Semantic Clusters of Identities

Our primary question for the case study was to understand how identities clustered into semantically coherent, "institutionalized" sets of identities. In order to extract these sets, or clusters, of identities, we make two assumptions. First, we assume that the use of each identity is a "mixture" over a finite set of latent identity clusters. This assumption is generally supported by social theory - recall, for example, that MacKinnon and Heise emphasize the differential association of identities to latent institutions they extract from a semantic network analysis (Heise, 2010a). Similarly, we assume that individuals are a mixture over institutions. This assumption is based on literature which suggests that social processes, like segregation, create disparities in the social contexts that we frequent. The social contexts in which we reside in turn influence our perceptions of the identities of those around us (Tajfel and Turner, 1979; Turner et al., 1987).

In defining both people and identities as mixtures over latent, institutionalized sets of identities, a natural algorithmic fit to extract these latent identity sets is LDA (Blei et al., 2003). We apply LDA to the user by identity matrix M, where each cell of the matrix $M_{u,i}$ represents the number of times a particular user u mentioned the identity i. To avoid issues with shorter documents, we only run the LDA on users with more than 50 unique identity labels. To avoid reliance on universally common terms, we drop identities tweeted by more than 50% of our users, and to address sparsity we drop terms tweeted by fewer than 100 users. After cleaning, we are left with 161K users (65% of the original set) and 4293 identities (approximately 5% of the full set captured).

The number of topics for LDA were set based on domain knowledge. In particular, as Heise and MacKinnon observed only twelve taxonomic collections of identities in their qualitative analysis of WordNet, we kept the number of topics k to a lower number (30) than is traditional in the topic modeling literature. We use the version of LDA implemented in gensim (Rehurek and Sojka, 2011) and allow the concentration parameter α to be estimated from the data.

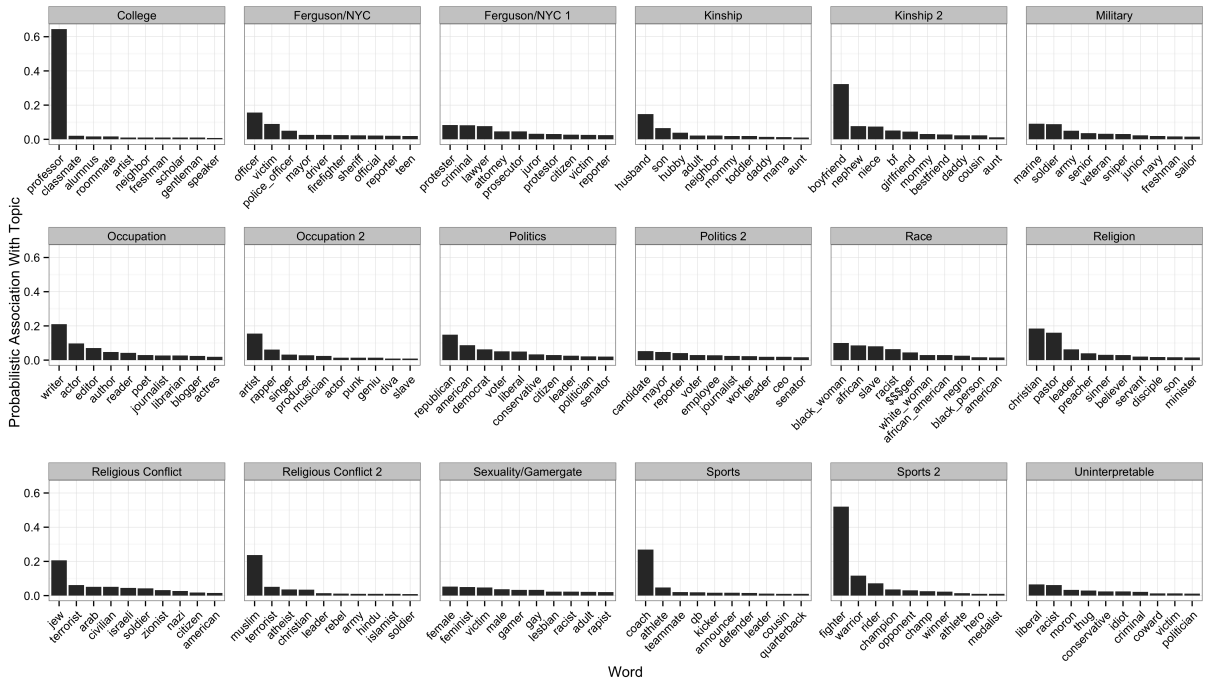


Figure 3.4: Results of the LDA. Each sub-plot is an interpretable topic. Within each plot we show the top 10 words associated with the topic. Bar height represents the probabilistic association of the identity to the identity cluster based on the posterior of the model

Figure 3.4 shows one plot for each of the 17 identity clusters we were able to interpret from the LDA, along with one example uninterpretable topic that was typical of the 13 topics we could not provide a coherent label for. Labels for each cluster, provided by us, are given in the grey headers for each subplot. Within each subplot, we show the top ten identities for that identity set, along with the identities' associations to the topic as defined by the posterior from the LDA model.

Of the twelve identity taxonomies that MacKinnon and Heise identified from their qualitative study of WordNet, we are easily able to observe seven of them from the clusters extracted by our model. At least one and sometimes two topics were observed that closely identified with the political, kinship, religion, race/ethnicity, leisure/sporting, occupation and sexuality classifications they provided. This connection to prior work gives us confidence that our approach can reproduce traditional, denotive, taxonomic clusterings of identities. In addition to those taxonomic clusters found by MacKinnon and Heise in WordNet, we also find strong evidence for two addition clusters that fit this description, namely the school/college and military taxonomies of identities. These additional categories show the value of exploring semantic patterns in larger datasets using unsupervised approaches.

One advantage of Twitter in particular as a data source is that a small portion of tweets contain geospatial information. This allows us the opportunity to observe how spatial indicators of social context might influence the use of particular identities clusters by particular individuals. As an initial exploration of this, we consider how use of the "Race" identity cluster changes based on

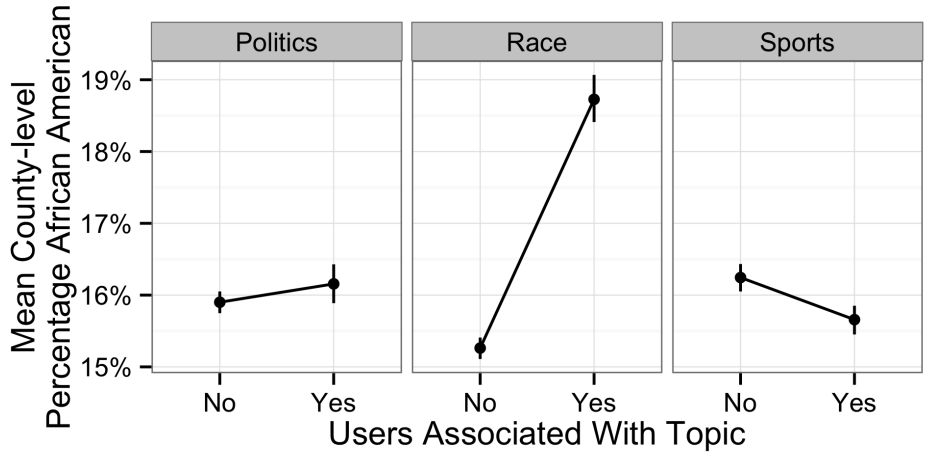


Figure 3.5: Differences in racial make up of geotagged users' counties for three identity clusters. The x-axis differentiates users who were associated and not associated with each cluster. The y-axis shows the percentage of the users' county that was African American. Error bars are 99% bootstrapped CIs.

the racial make-up in the area a user tweets from most often.

From our data, we extract 71K who have at least one geotagged tweet from within the United States. For each user, we determine the county within which they tweet most frequently, and then retrieve information from the 2013 American Community Survey on the percentage of that county's residents who are African American. As the posterior distributions of user to identity sets was heavily bimodal, we discretize user associations to each identity cluster into a binary variable. Each geo-tagged user is thus represented here by the percentage of African Americans in their county and a set of binary variables representing whether or not she used is associated with each of the 18 identity clusters in Figure 3.4.

In terms of the expected relationship between context and use of identities in the Race cluster, two competing hypotheses are relevant. First, the *contact hypothesis* (Pettigrew and Tropp, 2008), perhaps the most well-tested social psychological theory, states that the more frequently we come in contact with someone of a particular race, the more favorable our view of that race will be. Given the negative connotation of several of the identities in this cluster (e.g., "$$$ger","slave", "negro"), we might thus expect that use of this cluster of identities is associated with a lower African American population in the users' county. In contrast, more general models of associative cognition (e.g., (Anderson, 2007)) suggests that regardless of affect, the more we come in contact with someone having a particular identity, the more likely it is we will think about and talk about that identity and identities semantically associated related to it.

Figure 3.5 shows 99% bootstrapped confidence intervals for the average percentage of African Americans in a users' county for users that were and were not associated with three identity clusters - race, politics and sports. The latter two are included simply as points of comparison. We find that in places where race enters the social context to a greater degree-that is, in places where the percentage of African Americans is greater, people are more likely to use racialized identities. In contrast, we see in Figure 3.5 that there is little, if any, practical difference between the use of identity labels in the "Politics" and "Sports" identity sets according to race.

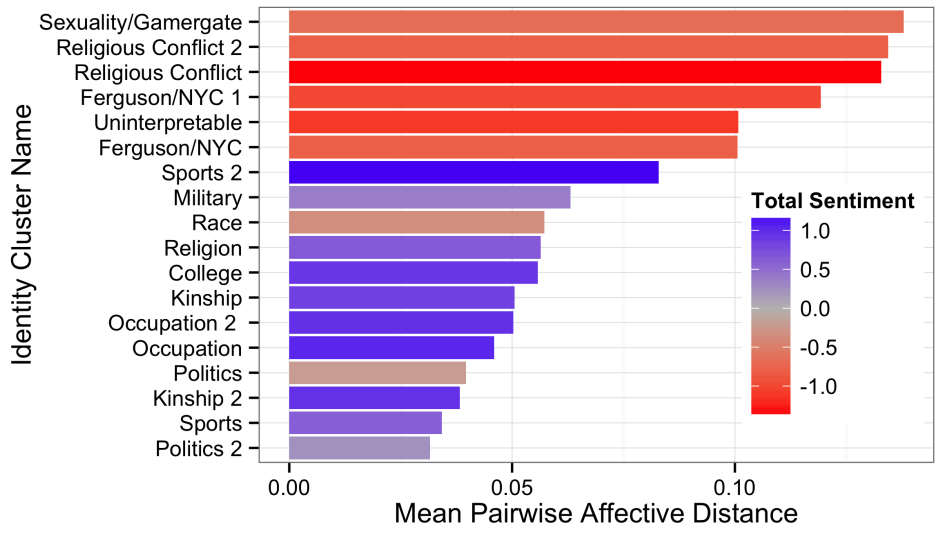


Figure 3.6: Affective meanings of the different identity clusters. Color indicates the sum of the affective meanings for each cluster - the darker the color, the more negative the affect. The size of the bar represents the mean pairwise distance between affective meanings of identities within the cluster

This finding by no means attempts to discredit contact theory, rather, it simply suggests that the denotive, purely semantic coherence of the Race identity cluster may be stronger than its affective coherence. Consequently, general theories of associative cognition are more applicable than theories which focus on affective relations across racial lines. Semantic coherence also seems to be important for the several identity clusters in our data that are specific to particular social issues. For example, two clusters are relevant to religious conflict, containing terms relating to both military and religious identities. The Sexuality cluster also contains the term "gamer", a connection to the Gamergate controversy that has led to frank discussions about sexism in the video gaming culture. Finally, as we would expect, we find two clusters relating to the events that occurred in Ferguson and New York City.

As these identity clusters cover emotionally charged and ever-evolving social issues, we would expect them to have a very different set of *affective* meanings than the more static and less emotional institutionalized topics found by MacKinnon and Heise in WordNet and replicated in our analysis. This assumption can be tested empirically by comparing affective meanings across the different identity clusters. In order to test this assumption, we take a coarse-grained measure of affective meaning for each identity in Figure 3.4. We use Vader (Hutto and Gilbert, 2014), a lexicon-based, tweet-level sentiment analyzer and apply it to each tweet containing an identity. The "affect score" for each identity is the average sentiment of each tweet that it is seen in. Importantly, before running Vader we remove from the sentiment lexicon all identity terms.

Figure 3.6 presents two pieces of information about the affective nature of each identity cluster. First, along the horizontal axis, we plot the mean pairwise distance of the affect scores for the identities that represent each cluster. The higher this value, the less similar the affective meanings of the identities within the cluster are. Second, the color of each bar represents the sum of the affective scores for the identities within the cluster. Here, the darker the color (i.e.,

the closer to black the bar is), the more negative the identities were in total, and the lighter the color, the more positive they were.

Patterns in Figure 3.6 support our prior suggestion that identity clusters which develop around social issues have stronger affective contrasts. As we would expect, they carried significantly more negative emotion as well. These clusters are interesting in that they are semantically coherent but affectively disparate. While work remains to be done, our ability to uncover identity clusters fitting this description may help to learn more about how affective and semantic meanings co-evolve during complex social events, like those that occurred in Ferguson and New York City last year.

3.7 Case Study - Arab Spring Twitter Data

3.7.1 Overview

In addition to the above case study presented in the WWW paper, this document also contains a case study using the same analysis approach but on a dataset of 156K Twitter users who frequently tweeted about the Arab Spring. The data for this set of users were collected specifically for this case study and were collected in the following way.

I first obtained Twitter data collected from two sources from April 2009 to November 2013 that has been used for other related work by CASOS (e.g. Wei et al., 2015a). The first source was collected by tracking a manually curated set of keywords, users and geo-boxes related to the Arab Spring using the Twitter Streaming API, which returns a maximum of around 1% of the full set of tweets at any given time. Parameters used to search the Streaming API focused mostly on events surrounding Egypt, Libya, Syria, Tunisia and Yemen, though certain parameters did apply to the entire region associated with the Arab Spring. The second way in which tweets were obtained was from an outside researcher who provided us with geo-tagged tweets from a 10% sample of the full set of tweets during this same time period. These geo-tagged tweets were only obtained from the set of countries given in Figure 2.1.

This dataset consisted of approximately 81M tweets. For this case study, I obtained a filtered dataset by selecting out tweets written in English and sent by users who tweeted more than five times and were active for at least a week in the original dataset. In total, this set constituted around 11.2M tweets from around 300,000 users. From here, I then obtained the last 3200 tweets of 156K of these Twitter users using the same method described in Section 3.3.1.

After this data collection, I then ran an identity classifier on tweets from these 156K users that fit the qualifications provided in Section 3.3.1 (User had 50-15K tweets and ¡25K followers, no retweets, tweets with URLs or non-English tweets). This classifier was a slightly updated version of the prior classifier, which was more strict on its classification of the word “man” as an identity in certain cases. This behavior was modified by using an altered training dataset where all instances of the phrase “man” were labeled as identities (in the original dataset, “man” appeared in the phrase “Oh man” in several training examples and was not classified as an identity). It seems, qualitatively, that this later version is slightly preferable to the former, but other than this change, the classifier was the exact same as the one utilized in the Ferguson case study.

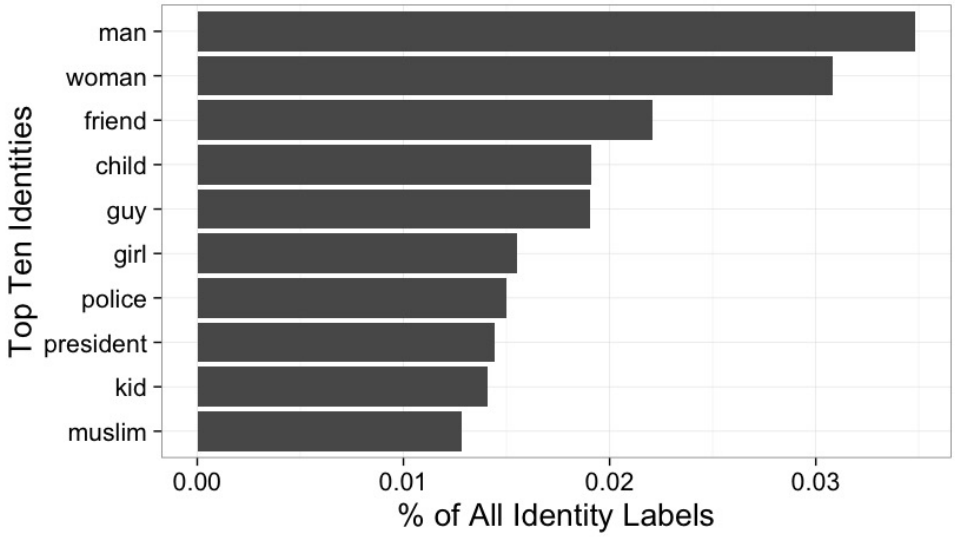


Figure 3.7: Top ten identities in the Arab Spring dataset

3.7.2 Results

Figure 3.7 shows the top ten identities from users in the Arab Spring dataset. As expected, this version of the classifier labels "man" as the most frequent identity, a contrast from what we see in Figure 3.2. Beyond this technical difference, we see that six out of the top ten identities in the Ferguson dataset (girl, guy, friend, kid, woman, child) are in the top ten in the Arab Spring data as well. These commonalities suggest that these standard gender and age-based classifications are relatively universal.

Figure 3.7 also shows that beyond these universal identities, context can have strong impacts on results. The identities "fan", "dude" and "ni$$a" appear only in the top ten identities of the Ferguson data, while the identities "president", "muslim" and "police" appear only in the top ten identities of the Arab Spring data. The identities "ni$$a" and "Muslim" are strongly reflective of the different populations from which users were drawn in the two different case studies, showing that even a very quick summarization of the important identities in a dataset can reveal important focii of the population of interest. This provides evidence that identity surveys can be useful as a summarization mechanism of a particular dataset in addition to summarizing different topical interests within a dataset.

Figure 3.8, which presents the distribution of unique identities per user in the Arab Spring data, provides further evidence that individuals on Twitter use a diverse array of unique identity labels when interacting online. In fact, individuals in the Arab Spring data used considerably more unique identities - a median of 104 - than users in the Ferguson data. This fairly high number of identities gives us further confidence that users are discussing identities across multiple social settings, which means that the use of LDA is probably an appropriate method to extract clusters of identities.

We therefore engage in the same process used above of treating each user as a multinomial over identities and running LDA on this matrix. Fifteen of the most interesting identity "topics" uncovered are displayed in Figure 3.9. Again, we see both important commonalities and impor-

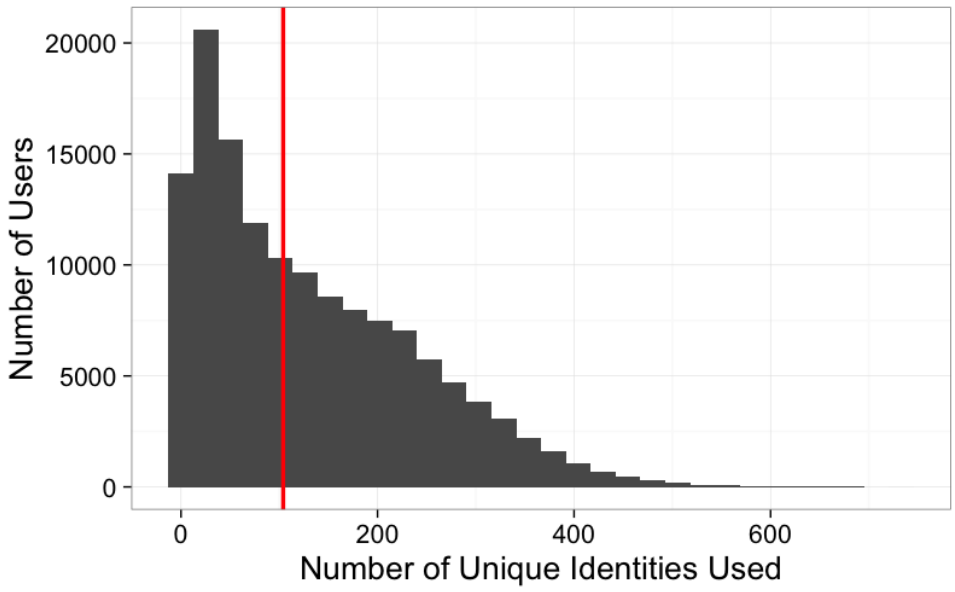


Figure 3.8: Distribution of unique identities used by Arab Spring users

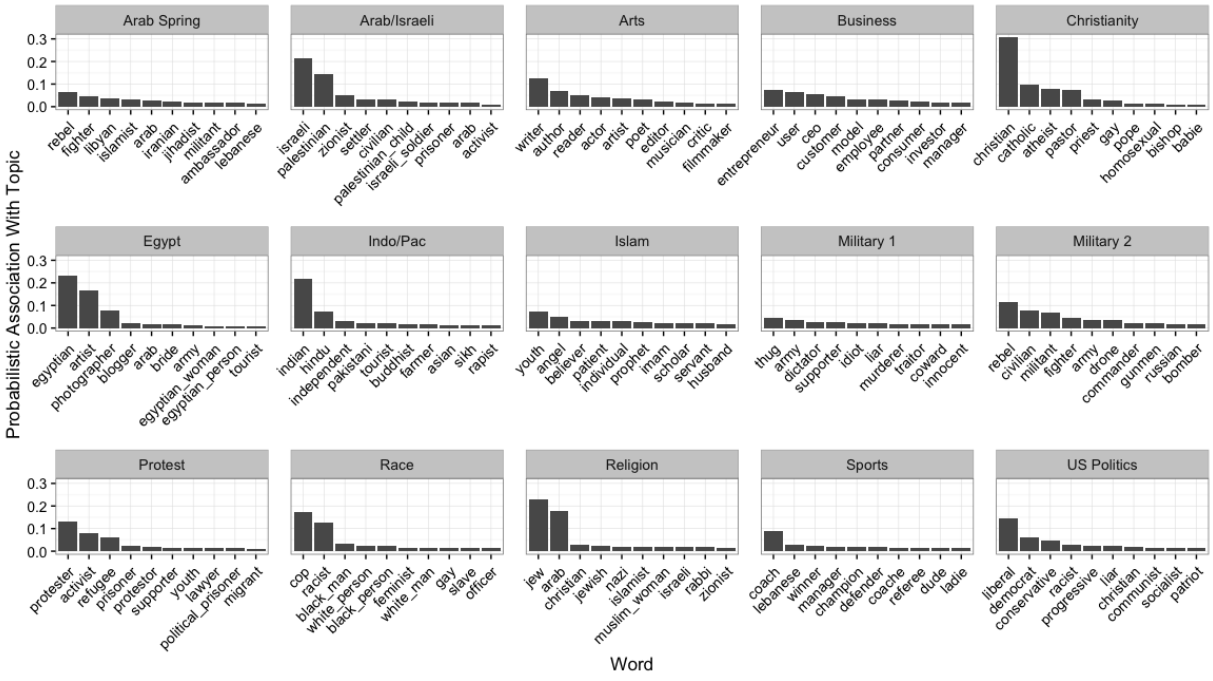


Figure 3.9: Named clusters extracted from the Arab Spring data

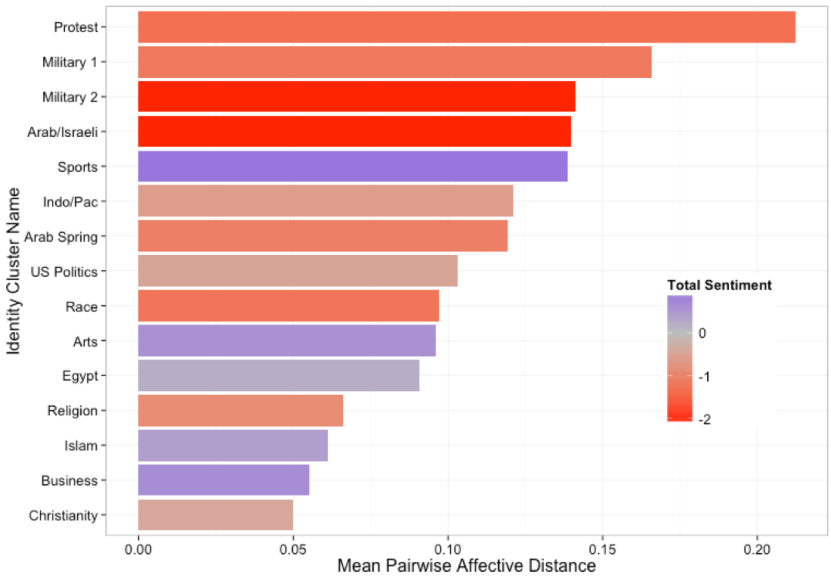


Figure 3.10: Affective meanings of the different identity clusters. Color indicates the sum of the affective meanings for each cluster - the darker the color, the more negative the affect. The size of the bar represents the mean pairwise distance between affective meanings of identities within the cluster

tant differences between the Arab Spring data and the Ferguson data. In terms of commonalities, fairly static "institutions" are uncovered - common topics across the two datasets include sports, politics, religion, race and the military. However, in contrast to the Ferguson data, we see in the Arab Spring data clusters of identities related specifically to Islam and Christianity, as well as to events in Egypt and, of course, to the Arab Spring.

Combined, commonalities between Figure 3.9 and Figure 3.4 present evidence in favor of Heise and MacKinnon's (2010) notion that static, persistent institutions influence how we organize the world. However, they also present evidence that *emergent institutions*, which I take to include both short social events (e.g. protests) and prolonged social events (e.g. the Arab Spring) create important organizations of identities as well. How, exactly, these emergent institutions persist or dissipate is perhaps an interesting question for future work. Inherently, this observation seems similar to Smith-Lovin and Douglas's (1992) discussion of how boundaries may form across social groups with the development of different affective meanings of particular identities. Perhaps this view can be amended to consider the possibility that boundaries between social groups may arise with distinct affective meanings of particular identities, and institutions may arise out of reorganizations of semantic associations between identities.

The final element of the Ferguson case study to be replicated is the study of affective meanings of identities within the different "identity topics" uncovered. Figure 3.10 replicates Figure 3.6 with the Arab Spring dataset. Figure 3.10 suggests a broad agreement with conclusions made above - topics relevant to conflict were generally negative and fairly disparate, in particular those related to protests. However, support for this supposition is not as clear-cut as in the Ferguson data - for example, a topic we judged to be generally relative to the military was both highly negative and had a high mean pairwise affective distance. This could be due to the heuris-

tic method used to label topics, but may also be due to the heuristic approach used to measure affective meaning of identities. I address the latter point in Chapter 4, the former point is an interesting case to consider in future work.

3.8 Conclusion

The present work makes two major contributions to the literature on identities and their use in text. First, we pose the novel problem of extracting all identities from a given corpora and develop a straightforward machine learning approach to address this problem. Our model outperforms a strong rule-based baseline by 33% on a validation set. Along the way, we develop new standards for determining what constitutes an identity label in text and produce a new, public, labeled corpora for other researchers to utilize.

Second, we perform a case study using our classifier on a large corpora of Twitter data collected around the Eric Garner and Michael Brown tragedies. We analyze semantically coherent clusters of identities and find they have important connections with a previous study on such structures (Heise and MacKinnon, 2010). We also observe identity clusters that are affectively disparate and highly negative but that are still semantically cohesive. A closer evaluation of the temporal and spatial dynamics of identities in these clusters, particularly those relating to the Eric Garner and Michael Brown tragedies, may provide unique theoretical insights into how semantic relationships between identities coevolve with their affective meanings. Finally, we consider how social contexts of Twitter users affect their use of particular identity labels. Specifically, we observe a positive association between the percentage of people in an individuals' home county that are African American and her use of racial identities on Twitter.

While these findings are fairly general and are supported by existing literature, conclusions should be taken with care for several reasons. First, we focus on a particular domain using a particular source of data. Second, the classifier we develop could be improved in several ways. In addition to issues stated in our error analysis, it is likely that we can make better use of our unlabeled data than by simply constructing a dictionary. Though we did not investigate this in the present work, it is possible that using the unlabeled data in this way may have introduced significant noise into the feature space. We also should be able to leverage information in knowledge bases to improve classification and to perform Word Sense Disambiguation.

Regardless of these drawbacks, our case study findings present several interesting avenues of future work that can already be addressed with the tools developed here. For example, we have not yet considered how identities cluster on purely affective meaning as opposed to on semantic connections. Further, we have not perform any sort of temporal analysis. Finally, as opposed to affective analysis at the tweet-level, concept level analysis of sentiment would likely prove interesting in understanding how different people have different feelings towards the same identities.

3.9 Acknowledgements

We would like to thank the Data Science for Social Good fellows who volunteered their time for this article. We would also especially like to thank Jonathan Morgan, David Bamman and Waleed Ammar for helpful discussions.

Chapter 4: Extracting Affective and Semantic Stereotypes from Twitter

A social identity is a word or phrase used to define a particular type, group or class of individuals (Smith-Lovin, 2007). The identities we choose for ourselves or that are chosen for us impact our lives in important ways. For example, being labeled a woman or an African American can have a significant negative effect on your employment opportunities (Bertrand and Mullainathan, 2003).

How we are impacted by our identities is determined by the *stereotypes*, or meanings, attached to them. Stereotypes have been represented in both an *affective* way, i.e., how people feel about an identity (Fiske et al., 2002; Heise, 2007), and in a *semantic* way, i.e. what other identities an identity is associated with (Freeman and Ambady, 2011; Kunda and Thagard, 1996). Affective stereotypes and semantic stereotypes are often related - when we think about "heroes", we also are likely to think about "champions". These two identities are both affectively similar and semantically associated. However, affective and semantic stereotypes can also be quite distinct - "heroes" are semantically associated with "villains", but are not affectively similar to them. In a related vein, "president" and "dude" are highly affectively related, but are not semantically so.

Unfortunately, little work has been done to understand these (dis)connects between affective and semantic stereotypes. There are two primary reasons for this. First, the communities from which these two models have arisen are different- affective stereotype models have come out of social psychology, whereas semantic models have generally been restricted to cognitive psychology. Second, and more relevant to the present work, scholars have found it difficult to analyze affective and semantic stereotypes in a consistent fashion - cognitive psychologists generally rely on experimental procedures to capture semantic relationships, while social psychologists generally rely on surveys to capture affective meanings.

The present work introduces a Bayesian latent variable model to measure both the affective and semantic stereotypes of Twitter users. In doing so, we present a new conceptual model of stereotypes as both the affective meaning and semantic associations of an identity, as well as a way to measure both forms of stereotype in a consistent fashion. Our approach to extracting semantic associations between identities is inspired by cognitive semantic models of stereotype and draws on recent literature in efficient inference for Bayesian models of text (Chen et al., 2012). Our approach to extracting affective stereotypes extends prior work synthesizing modern natural language processing (NLP) tools (Ahothali and Hoey, 2015; Joseph et al., 2016b) with Affect Control Theory (ACT) (Heise, 2007), a prominent social psychological theory of affective stereotypes.

The rest of this article is structured as follows: we first provide a review of related literature

from a variety of disciplines. We then describe our model, how we assess its fit to the data, and its performance. Finally, we provide a case study of our model on 45K Twitter users who were active in the discussion over the deaths of Michael Brown and Eric Garner, as well as concluding thoughts. The case study covers interesting connections between semantic stereotypes and the sociological concept of *institutions* introduced by Heise and MacKinnon (2010). We then consider one particular identity, "thug", that is relevant to our dataset. We highlight the advantages of considering both affect and semantic stereotypes as well as how the meaning inferred by our model is similar to and different than meaning inferred using similar data by recent computational linguistic models (Pennington et al., 2014)

4.1 Literature Review

A variety of research has leveraged NLP techniques to study stereotyping. Bamman et al. (2014) extract profiles of literary character types using a bag-of-words representation. Fast et al. (Fast et al., 2016) study gender stereotypes in a large corpora of fiction, focusing on how different verbs and adjectives are associated differently with men as opposed to women, and how stereotyping affects the popularity of a particular story on the website Wattpad. Several researchers have also considered biases in Wikipedia articles by gender (Bamman and Smith, 2014; Graells-Garrido et al., 2015).

While these articles suggest the promise in utilizing NLP methods to study stereotyping, they are also somewhat disconnected from the social psychological literature on mathematical models of stereotypes. In the present work, our aim is to connect social psychological work in this area to existing NLP toolkits- the rest of this section explains where such connections have already arisen and how we will leverage them. In addition, it should be noted that there is also a larger literature on how stereotypes impact decisions on how to express information about identities (Beukeboom and others, 2014). Our work complements these efforts, as we focus on a different set of social psychological models that therefore offer a different view of how stereotypes manifest themselves in language.

4.1.1 Semantic Relations as Stereotypes

Our model of semantic stereotypes is based on the *parallel constraint satisfaction model* (here, PCS model for short), a cognitive psychological model where stereotypes are defined by positive and negative links between nodes arranged in a semantic network (Freeman and Ambady, 2011; Kunda and Thagard, 1996; Schröder and Thagard, 2014). In most PCS models, nodes represent identities and links represent cognitive associations between them. Cognitive activation spreads along positive links and is inhibited by negative links in a way that lead us to label people in certain ways. Links in PCS models are generally set manually by the researcher; even a simple tool to extract prominent semantic relationships from text would thus be useful for PCS modelers.

If one assumes a Gaussian distribution over activation at each node in a PCS model, they can alternatively be thought of as Gaussian Markov Random Fields (GMRF) with particular semantics on the links and nodes. GMRFs can themselves be represented via the inverse of a Gaussian covariance matrix; the problem of extracting the semantic network of a PCS model from text thus reduces to extracting a Gaussian covariance matrix from text that represents correlations in activation scores. Conveniently, the Correlated Topic Model (CTM) (Blei and Lafferty, 2007)

presents a method to do just that, except with correlations between topics instead of between words (identities) themselves. By removing the assumption of topics from the CTM and assuming Twitter users are "bags of identities", we will be able to obtain the desired covariance matrix, which will measure which pairs of identities tend to be used by similar sets of individuals.

Our model will thus explicitly measure the extent to which identities are associated (Deese, 1966; Hill et al., 2014b; Miller and Charles, 1991), or semantically related (Resnik, 1999). A significant amount of work has also focused recently on the use of "predict models" (Baroni et al., 2014), i.e. deep word embedding approaches (Levy and Goldberg, 2014; Mikolov et al., 2013a; Pennington et al., 2014) to capture words that are both associated and, preferrably, words that are semantically *similar*, i.e. words that are "substitutable" for one another (Sahlgren, 2006)[1]. While models of stereotypes have considered, under different names, both association and similarity (Freeman and Ambady, 2011; Kunda and Thagard, 1996; Schröder and Thagard, 2014), our focus here is explicitly on association. Such associations between identities have been uncovered using semantic network analysis of dictionaries by sociologists David Heise and Neil MacKinnon (2010), who refer to them as "institutions". For example, the "family institution" includes the identities "brother","mother", etc.

NLP researchers have rarely considered identities as a unit of study, but work on Fine-Grained Named Entity Recognition, which requires a set of labels for individuals found in text, has utilized existing databases of semantic relationships between identities (Del Corro et al., 2015; Ekbal et al., 2010). These scholars generally utilize Wordnet, which implies a hierarchical structure of semantic stereotypes (e.g. identities related to occupations, those only related to occupations at a hospital, etc.). Joseph et al.'s (2016b) work is an exception - they develop a method to extract identities from tweets and use LDA to uncover semantic clusters of identities. The authors find that several clusters do not seem to fit neatly into hierarchies as defined by WordNet, but do so via fairly ad hoc methods and do not pursue an exploration of this point as we do here. Bamman et al. (2014) also extract stereotypes from text, but focus on associating identities to all words rather than specifically to other identities.

4.1.2 Affective Meanings as Stereotypes

Affect Control Theory (ACT) is a social psychological model of affective stereotypes- for a comparison to other models and a review, see (Hoey et al., 2013b; Rogers et al., 2013). ACT assumes a particular measurement system for the affective stereotypes of identities, the behaviors these identities engage in, and modifiers (e.g. "bad") that can be used to describe identities. Affective meanings of these entities are defined in a three dimensional *EPA* space with axes entitled *E*valuation (goodness/badness), *P*otency (strength/weakness) and *A*ctivity (activeness/passiveness), each spanning the range of -4.3 to +4.3.

ACT also describes how *social events* change the stereotypes we hold. Social events can also be used, as we will show, to infer stereotypes from text. A social event is an interaction in which an *actor* identity enacts a *behavior* on an *object* identity (Heise, 2007). ACT's quantitative framework defines how EPA ratings of the actor, behavior and object change after observing the event. This model represents our intuition that, for example, anyone seen doing a "bad thing" to a "good person" must themselves be somewhat "bad".

[1]Though parameterizations can mediate this distinction (Hill et al., 2014b; Sahlgren, 2006)

Social events have a *pre-event transient* meaning, f, that is modified by the social event to produce a post-event transient meaning, f'. Both f and f' are vectors of length nine, one element each for the Evaluative, Potency and Activity sentiment dimensions for the $a$ctor, $b$ehavior and $o$bject :

$$f = \begin{bmatrix} a_e & a_p & a_a & b_e & b_p & b_a & o_e & o_p & o_a \end{bmatrix}$$

ACT provides a regression equation that is used to determine the elements of the post-event transient f' as a function of f. Mathematically, $f' = \mathcal{M}g(f)$, where $g(f)$ is a kx1 vector of covariates (e.g. $\begin{bmatrix} 1 & a_e & \dots & b_e o_e & \dots & a_e b_e o_e \end{bmatrix}$) and $\mathcal{M}$ is a 9xk matrix specifying 9 different sets of regression coefficients, one for each element of f'. Importantly, $g(f)$ consists of only linear combinations of the elements of f. The actual covariates $g(f)$ and coefficients $\mathcal{M}$ used in this model are estimated via survey data; we refer the reader to (Morgan et al., 2015) for details. Assuming the form of $g(f)$ and the values in $\mathcal{M}$ are given, as we do here, the post-event transient can simply be constructed as follows, where $\mathcal{M}_x$ represents row x of the coefficient matrix:

$$f' = \begin{bmatrix} \mathcal{M}_{a_e} \cdot g(f) & \mathcal{M}_{a_p} \cdot g(f) & \dots & \mathcal{M}_{o_a} \cdot g(f) \end{bmatrix}$$

ACT also allows for a regression model in which modifiers change how social events impact perception by changing the meaning of an the actor or object (e.g. a "bad teacher" is different than a "teacher") - for details, we refer the reader to (Heise, 2007).

Given f and f', we can compute the *deflection* of a social event as the unnormalized Euclidean distance between the pre-event and post-event transients:

$$deflection = \sum_j^9 (f_j - f'_j)^2 = \sum_j^9 (f_j - \mathcal{M}_j \cdot g(f))^2 \tag{4.1}$$

Deflection gives an idea of how "expected" a social event was - high deflection means the affective meanings have changed dramatically, signifying an unexpected occurrence. The deflection equation is vital to our use of ACT as a model of affective stereotypes because it can also be used to determine the optimal (in the sense that the event becomes "most expected") EPA profile of the actor, behavior or object given information on the other two. In this way, social events that are discussed by an individual on Twitter can be used to infer the individual's stereotypes.

A key observation in the present work is that there is a straightforward way to include additional factors into the deflection model. While previous work using ACT for text data has relied solely on the existing social event framework (Ahothali and Hoey, 2015; Joseph et al., 2016b), the sentiment analysis literature contains, of course a vast array of additional constraining factors on sentimental meaning of words in text. Here, we introduce a framework for including these additional factors into the deflection model.

Sentiment mining of Twitter is a popular area of research (Hutto and Gilbert, 2014; Liu et al., 2012; Rosenthal et al., 2015; Taddy, 2013). Although most of this work focuses on analyzing sentiment held in a particular tweet, such approaches can be used to construct a rough measure of sentiment held towards particular entities within tweets (Chambers et al., 2013), and we utilize a similar approach as a baseline. Approaches assessing the affective meaning of a concept or expression across an entire corpus of Twitter data have also been developed (Chen et al., 2012;

Kiritchenko et al., 2014). Researchers have also considered the problem of *target-dependent* sentiment analysis for Twitter data (Jiang et al., 2011), which focuses on whether or not a particular tweet (as opposed to the entire corpus) is positive or negatively focused on a particular concept. Our model to extract the affective component of stereotypes is similar to several of these efforts in our use of a dependency parse to extract additional sentiment information from the text (Dong et al., 2014; Zhang et al., 2015).

4.1.3 Combining Affect and Semantics

Computational linguists have also begun to develop approaches to jointly consider both semantic and affective meanings of words (Maas et al., 2011; Tang et al., 2014; Vo and Zhang, 2015). Perhaps most relevant, recent work has developed *sentiment-specific* word embeddings from Twitter data (Tang et al., 2014; Vo and Zhang, 2015). The focus in these models is generally on constructing continuously-valued vectors of word meanings that are constructed based on *both* semantic and affective meaning, rather than modeling these two concepts separately as we do here.

Yang and Eisenstein (2015) learn sentiment-based embeddings that are allowed to vary by social community as well. Interestingly, Affect Control theorists have generally found that affective stereotypes of identities are relatively consistent across most social boundaries (Heise, 2007; Robinson et al., 2006). Similarly, cognitive psychologists generally assume semantic links between identities to be relatively consistent (Freeman and Ambady, 2011). In the present work we assume no parametric differences across social boundaries, for these reasons and because we are already subsampling from a fairly specific set of individuals interested in a particular topic.

4.2 Data

The Twitter dataset we use is a collection of all tweets from 2013-2015 of 44,896 users. Data was originally collected through the Streaming API from August through December of 2014, using keywords that were focused on the events surrounding the deaths of Eric Garner and Michael Brown. Once data collection was finished in December, 2014, we then selected all users who had sent five or more tweets captured by our keywords and gathered the last 3200 tweets they had sent.

In October 2015, we re-collected data for all 250K users who sent at least one geotagged tweet.[2] The present work focuses on a subset of these users for which we were able to obtain all tweets from 2013-2015, who sent between 250 and 10K tweets, had less than 50K followers, and who had at least 50 tweets that contained one or more of the 310 *identities of interest* studied in the present work. We did not consider retweets or tweets with fewer than five unique non-punctuation terms.

Note that the identities of interest in the present work do not encompass all identities used by Twitter users - this particular set was selected based on frequency of use within our dataset as well as from identities of interest to prior research (Davenport, 2016; Joseph et al., 2016b). In addition to the Twitter dataset and identity list described, we also leverage two large dictionaries of words matched to their EPA profiles (Smith-Lovin and Robinson, 2015; Warriner et al., 2013),

[2]Geotags were not used in the present work but were instead relevant to concurrent efforts for which we wished to have a consistent sample

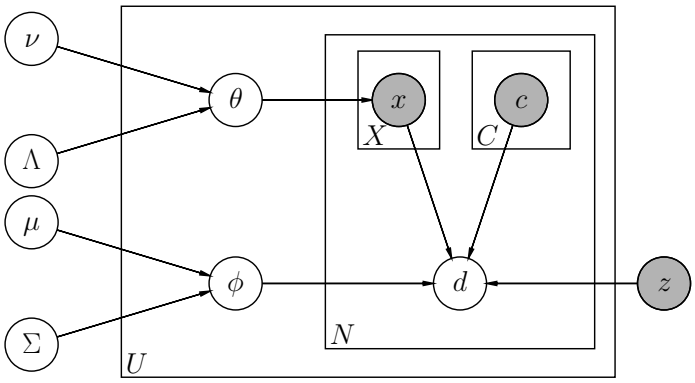


Figure 4.1: A graphical model of our method.

collected via survey methods. We also leverage EMOLEX (Mohammad and Turney, 2010), a survey linking identities to emotions. The emotions words are transformed into EPA values by mapping these emotion words into our two EPA dictionaries. Finally, we utilize the Emoji dataset from (Kralj Novak et al., 2015) to leverage sentiment expressed via these tokens. All of these datasets are publicly available; code and data from the present work will be released on the first author's Github page (`http://www.github.com/kennyjoseph/tacl_pub`).

4.3 Model

The statistical model we use to infer semantic and affective stereotypes is presented in Figure 4.1 (without hyperparameters). For each user u in our dataset U, we have N_u tweets. Each tweet n for user u contains a (possibly empty) set of identities of interest found in the tweet's text, $X_{u,n}$. We consider any noun or adjective whose surface form is in our set of identities of interest to be an identity. If none are found, we ignore the tweet.

Each tweet may also contain a set of "constraining words", $C_{u,n}$. Constraining words are words in a tweet that are in our EPA dictionaries but that are not identities of interest (e.g. behaviors). We construct $C_{u,n}$ by finding all words in the tweet also in our EPA dictionaries. The EPA values of these words are held in z, which we will assume to be known and fixed. Thus, z_{w_e} gives the $e$valuative dimension of constraint word w. For example, in the tweet "all girls rule, all boys drool", the set $X_{u,n}$ would be comprised of $(girl,\ boy)$, $C_{u,n}$ would be comprised of $(drool,\ rule)$, as both of these are in our EPA dictionaries, and the word "all" would be ignored, as it is neither an identity of interest nor in our EPA dictionaries.

The model has two components. The first, consisting of parameters ν, Λ and θ, is used to infer semantic relationships between identities. The parameter matrix θ estimates the extent to which a particular user uses a particular identity, θ_u is a row of this matrix defining values for the user u. Following Eisenstein and colleagues (2014), we refer to values in θ as "activation scores". The parameters ν and Λ define mean and covariance parameters over these activation scores, respectively. Parameterization follows the CTM, with a logistic normal prior over θ.

The parameter ν is thus a vector of length $|I|$, where I is the set of identities of interest. The

parameter Λ is an $|I|x|I|$ covariance matrix of these activations, and θ is an $|U|x|I|$ matrix. We perform a fully Bayesian analysis, putting a conjugate Normal Inverse-Wishart prior over Λ and ν. Formally, this portion of the model can be defined as follows, where x is an element of $X_{u,n}$:

$$\begin{aligned} p(\nu, \Lambda) &\sim \mathcal{NIW}(\nu_0, \Lambda_0, \kappa_{0,A}, \gamma_{0,A}) \\ p(\theta) &\sim \mathcal{N}(\nu, \Lambda) \\ p(x) &\sim \text{Mult}(\text{softmax}(\theta_u)) \end{aligned}$$

The sentiment-based component of our model is parameterized as follows:

$$\begin{aligned} p(\mu, \Sigma) &\sim \mathcal{NIW}(\mu_0, \Sigma_0, \kappa_{0,S}, \gamma_{0,S}) \\ p(\phi) &\sim \mathcal{N}(\mu, \Sigma) \\ p(d) &\sim \text{Laplace}(q_{u,n}(\phi_u, X_{u,n}, C_{u,n}, z), \beta) \end{aligned}$$

Here, μ provides the mean affective ratings for each sentiment dimension of each identity and Σ provides the associated covariance matrix. Thus, μ is a vector of length $3|I|$, and Σ is a $3|I|$x$3|I|$ matrix. The parameter matrix ϕ gives per-user values for each sentiment dimension for each user; ϕ_{u,i_e} represents the element of ϕ corresponding to the $e$valuative dimension of the ith identity for user u. The core of the sentiment model is the development of the probability distribution of d.[3] This variable represents the sentimental (affective) *d*eflection of a particular tweet. Subscripts are given above for q, X, ϕ and C to emphasize that the parameterization of d is unique for each tweet and independent across users.

The variance, β, of the Laplace distribution over d is assumed fixed. Thus parameterization of $p(d)$ relies only on a mean function $q_{u,n}(\phi_u, X_{u,n}, C_{u,n}, z)$. This function will provide information about *sentiment constraints* that are observed in a tweet through it's sentence structure and the elements of $X_{u,n}$ and $C_{u,n}$. These constraints are a more general model of social events that serve to explain how "expected" the tweet is given users' current affective stereotypes. We now introduce how we move from a tweet's text to $q_{u,n}$ through a deterministic algorithm that extracts identities of interest and sentiment constraints.

4.3.1 Sentiment Constraint Extraction

We construct $q_{u,n}$ by extracting four types of semantic constraints - "clause-level", "emoji", "social event" and "social action" constraints. In order for us to do so, two preprocessing steps are necessary. First, each tweet is dependency parsed using a Twitter-specific dependency parser (Kong et al., 2014). Second, each tweet is run through a classifier (Joseph et al., 2016b) to determine if any elements of C are also identities (that are not in our set of interest).[4] We can then proceed to extract sentiment constraints, and we describe each constraint below. Note that constraints are formulated as a quadratic function of elements of ϕ_u, this is important for efficient inference. The u subscript is made implicit in this section in order to ease notation. Similarly, we will use X to represent $X_{u,n}$ and C for $C_{u,n}$.

[3]Note that while Equation 4.1 is deterministic, we follow prior work (Hoey et al., 2013b; Joseph et al., 2016a) in the assumption that it is more appropriate to think of deflection as a random variable, influenced by other unknown factors.

[4]We do not use this classifier to extract identities of interest in order to mitigate any biases it may induce.

Clause-level constraints are constructed on a per-identity basis (i.e. for each element of X independently) by taking the maximum absolute sentiment value for all elements of C that have the same root in the dependency parse. A clause-level (as opposed to a sentence-level) approach is possible due to the Twitter-specific dependency parser we use (Kong et al., 2014), which allows for multiple roots. For an identity x in X, let us define cl as the set of elements in C having the same root as x. Then a clause-level constraint is defined as $(\phi_{x_e} - \max(z_{c_e}; c \in cl))^2 + (\phi_{x_p} - \max(z_{c_p}; c \in cl))^2 + (\phi_{x_a} - \max(z_{c_a}; c \in cl))^2$. If cl is empty, no clause-level constraint is added for that x. We chose the maximum, as opposed to any other aggregation operator, because it produced results in pilot runs of the model that better fit intuitions about affective meanings.

If an emoji in our dictionary is found in a tweet, a constraint is added to the evaluative dimension of each identity in X with the affective value for that emoji. For each element of X, a constraint of the form $(\phi_{x_e} - z_{j_e})^2$, where j is the emoji's index in z, is added.

We then extract social event constraints. Our approach follows typical extraction of Subject, Verb, Object triplets using a dependency parse - we look for verbs that are in C that have a direct subject and object which are both identities, at least one of which is in X and both of which are in X or C. We also extract modifier terms, elements of C which are adjectives and direct descendants in the dependency parse of these identities, and apply the ACT modifier equation in these cases. Where multiple modifiers exist, values are averaged. Also, behaviors may have multiple terms, e.g. in the statement "the teachers were *forced to instruct* the students". In this case, we simply to average over all verbs constituting the action relationship between two identities.

Once an event is extracted, we introduce a social event constraint for that event. To introduce the mathematical form of a social event constraint, let us assume that an identity of interest x_k is found to be in a social event with identity of interest x_j where behavior c_b is enacted. We will first define the pre-event transient as follows, where m is the modifier equation that may incorporate values from C or from X:

$$f = [m(\phi_{x_{k,e}}) \quad m(\phi_{x_{k,p}}) \quad m(\phi_{x_{k,a}}) \quad z_{c_{b,e}} z_{c_{b,p}} \quad z_{c_{b,a}} \quad m(\phi_{x_{j,e}}) \quad m(\phi_{x_{j,p}}) \quad m(\phi_{x_{j,a}})]$$

Given the form of this pre-event transient, the full social event constraint can then be specified as in Equation 4.1, which defines the ACT notion of deflection.

Finally, we also extract social action constraints, which are social events in which no *object* can be found, but for which an actor in X and a behavior in C exist. Here, we just replace the EPA profile of the object with all zeros and construct a social event constraint with this f as above.

The mean function of $p(d)$ for a given tweet, $q_{u,n}(\phi_u, X_{u,n}, C_{u,n}, z)$, is the sum over all constraints uncovered in that tweet. Because it is a summation over constraints that are themselves quadratic in elements of ϕ_u, $q_{u,n}(\phi_u, X_{u,n}, C_{u,n}, z)$ is also quadratic in elements of ϕ_u.

4.3.2 Inference

Model inference is performed via Gibbs Sampling. We train on 85% of each user's tweets, retaining a 15% sample for evaluation. The Gibbs sampler can be split into two parallelizable components, one which infers parameters for the semantic model and one which infers parameters for the affective model. The Gibbs sampler for the semantic model is derived directly from

the work of Chen et al. (2013), who develop an auxiliary variable method to perform Gibbs sampling on the CTM. Beyond the removal of topics from their model, the inference procedure is identical and is thus not covered here due to space constraints.

The Gibbs sampler for the sentiment portion of the model is largely straightforward given previous results as well. Sampling for μ and Σ follows standard conjugate updates for the Normal-Inverse Wishart prior on the multivariate Gaussian distribution. The derivation for the conditional sampling distribution for each element of ϕ is the same, we choose ϕ_{u,i_e} as an example. For convenience, the u subscript is dropped from all variables below. Given all other variables (expressed as $\cdot$ below), the conditional sampling distribution can be expressed as follows:

$$p(\phi_{i_e}|\cdot) = p(\phi_{i_e}|\mu, \Sigma, \phi_{\neg i_e}) \prod_{n}^{N_i} p(d_n|\phi, \cdot) \tag{4.2}$$

We address the prior (left term on the right-hand side) and likelihood (right-most term) portions of Equation 4.2 separately. The prior requires that we condition over all other elements of $\phi_{\neg i_e}$ - through standard manipulations of the multivariate Gaussian distribution, we know that doing so leaves us with the following:

$$p(\phi_{i_e}|\mu, \Sigma, \phi_{\neg i_e}) \sim \mathcal{N}\left(\mu_{i_e} - \Sigma^{-1}_{i_e,i_e}\Sigma^{-1}_{i_e,\neg i_e}(\phi_{\neg i_e} - \mu_{\neg i_e}), \Sigma^{-1}_{i_e,i_e}\right) \tag{4.3}$$

The likelihood is also Gaussian, though the derivation requires slightly more thought. Recall that $p(d)$ is Laplace distributed. Let us assume, as we will throughout, that $\beta = 1$. As noted above, $q_{u,n}(\phi_u, X_{u,n}, C_{u,n}, z)$ deterministically returns an equation that represents quadratic constraints on elements of ϕ. Given values for all parameters except ϕ_{i_e}, and noting we are only interested in a sampling distribution for this particular element, we can ignore the value of d_n and say, for a particular tweet from a particular user, that:

$$p(d_n|\phi, .) \propto \exp(\frac{-|q_n(\phi, X_n, C_n, z)|}{2}) \tag{4.4}$$

Recalling that $q_n(\phi, X_n, C_n, z)$ is quadratic in ϕ_{i_e} by construction, it should be clear that with some rearranging of variables and by addressing the absolute value, we are left with a Gaussian distribution in ϕ_{i_e}. A formal proof of this is given in early work on Bayesian versions of ACT (Joseph et al., 2016a). The conditional sampling distribution of ϕ_{i_e} thus amounts to a conjugate normal update, with the only distinction being that each $p(d_n)$ has a unique variance. This simplicity is due to our restriction of the function q to be quadratic in elements of ϕ. Future work may consider alternative mechanisms to retain the theoretical ideals of ACT while still promoting efficient inference.

The Gibbs sampler for the sentiment portion of our model can therefore be run via iterative sampling of Gaussian distributions for each element of ϕ, followed by updates of μ and Σ. This algorithm is trivially parallelizable across users.

4.3.3 Model Performance Analysis

Our model, as well as the baselines, are evaluated based on their ability to predict which identities will appear in tweets in the test set. For the semantic component of the model, we evaluate performance using the probability-based metric of perplexity. For the affective portion, we chose

a ranking-based method to compare to the baseline, as such comparisons were more appropriate for reasons discussed below.

For the semantic portion of the model model, the log perplexity of the test data TD is measured as $\frac{-\sum_{u \in U} \sum_{n \in N_u} \sum_{x \in X} log(p(x))}{|TD|}$, where $|TD|$ stands for the total number of identities in all test tweets across all users. In theory, in order to calculate $p(x)$, the probability that identity x appears in $X_{n,u}$ (i.e. the odds of a particular identity in a particular test tweet for a given user), we would integrate out all latent variables in the model, equivalent in expectation to taking the average over many Gibbs samples of the perplexity metric. In practice, as with many other prior researchers (Chen et al., 2013; Griffiths and Steyvers, 2004; O'Connor et al., 2013), we find that variance across samples on the metric is low, and thus we average over only a few (5) Gibbs samples. We compare results with two baselines. The first is a multinomial over all counts for all users, and the second is a Laplace-smoothed language model. The latter is equivalent to a model where each user is defined by a multinomial distribution over the identities of interest with a symmetric Dirichlet prior.[5]

While computational linguists routinely evaluate sentiment models on gold-standard datasets, we take the sociological perspective of Dimmagio (2015), who argues that a true understanding of the affective meaning of a concept is not easily captured via human judgement. Instead, we choose to validate our model via the combination of predicting identities in tweets, a concrete task that can be derived directly from our data, and qualitative evaluation of the fitted sentiment profiles in our case study. More specifically in our case, there is little reason to believe that user stereotypes from this particular sample match our ACT survey data; we therefore do not use it as a validation or any annotator's judgements of their stereotypes.

There are three differences between our evaluation of the semantic and affective portions of our model. First, as opposed to considering perplexity, we compare the average rank of the correct identity as determined by our model versus the baselines. This decision is motivated by two factors - the lack of a truly comparable probabilistic baseline model for sentiment, and the fact that ACT's deflection model is typically used in the social sciences as a ranking model to compare the relative appropriateness of a few identities in particular social situations (Heise, 2007). Perhaps because of this latter point, when transferred to a probabilistic setting, the deflection model produces highly skewed predictive distributions, placing a unreasonable amount of certainty in the first few identities, leading the model to be a poor fit for perplexity-based problems (Joseph et al., 2016a). Qualitative comparisons we made within different model parameterizations thus suggested that differences in perplexity seemed to be based solely on the spread of the predictive distribution rather than spread rather than a model that was a stronger fit to the data.

The second difference between evaluation for the affective and semantic components is that as opposed to asking the model to rank on the distribution $p(x)$, we instead consider $p(x_k|C, X_{\neg k}, z)$. We therefore are evaluating the affective portion of the model's ability to make predictions given all affective information and all other identities in the tweet, a better way of evaluating the model's ability to use this information.

The final difference is in the baselines that are used. To develop our baseline models, we first run all training data through the VADER sentiment analysis tool (Hutto and Gilbert, 2014), which gives a continuous value on the interval [-1,1] for each tweet. We then compute, both for

[5]Choice of prior did not significantly effect results.

Semantic	Ppl.	Affective	Avg. Rank
Simple	4.864	Simple	134.744
User Baseline	4.474	User Baseline	127.272
Our Model	**4.363**	**Our Model**	**126.042**

Table 4.1: Results on the evaluation tasks. The left side of the table provides results for the semantic model and its baselines, the right side for the affective model and its baselines.

each user and overall, the average sentiment of all tweets in which each identity occurred. This value serves as the affective stereotype for the identity. For each identity in each test tweet, we compute the sentiment score for the tweet with that identity removed from the text.[6] We then compare this sentiment score to either to the vector of averages for the user (the *user* baseline) or to the average over all users (the *simple* baseline). For the user baseline, we use a simple fall-back model where the overall (simple) sentiment for an identity is used if the user has no tweets about a particular identity. Baseline models provide rankings based on which identities have affective stereotypes closest to the sentiment score for that test tweet.

4.3.4 Hyperparameters and Sampling

We set $\kappa_{0,A} = 100; \gamma_{0,A} = |I| + 1$. Parameters η_0 and Λ_0 were set to a vector of all zeros and the identity matrix, respectively. We leveraged survey data for μ_0, setting values for identities equal to their mean in the survey data and setting $\kappa_{0,S} = 300$. Where no survey data existed for an identity (3% of the cases), the prior was set empirically, by running a simple dictionary-based model over a random 1% of the training data. The parameter Σ_0 was the identity matrix and $\gamma_{0,S} = 3000$. Modifications to these priors did not greatly effect model performance.

The Gibbs sampler was run for 500 burn-in steps, as models converged quickly (little variation was observed after about 300 iterations). In the associative model, rapid convergence can be attributed to the large number of sub-iterations utilized by Chen et al. (2013) that we also utilize in our work, and to the fact that estimation was made much easier by the removal of topic structure. Rapid convergence in the affective model can be attributed to the relative lack of interdependency amongst parameters, which were intertwined only when event constraints were extracted from tweets. We took five samples for model evaluation, one every 100 iterations from the 500th-900th steps of the sampler.

4.4 Results

In this section, we provide results on model performance and then move to a case study.

4.4.1 Model Performance

Table 4.1 shows results on the prediction task developed to assess model fit, averaged across all five samples. Both portions of our model outperform their baselines, albeit by slim margins. The semantic portion of the model outperforms the user baseline by a relative increase of 2.5%, the affective portion by .97%. Although the goal of the present work is primarily to learn more about

[6] In order to retain sentence structure, we simply replace the actual identity with the word "identity"

stereotypes, the performance of the model relative to fairly strong baselines[7] gives us confidence that the parameters fit reasonably well to the data. Certainly, however, future work in a more predictive vein should be able to improve upon results.

Table 4.1 also shows that the semantic model component and its baselines significantly outperform the affective model and its baselines. This is to be expected - the affective portions of the model are given no information on the frequency with which users use each identity label, nor information on the extent to which each is used overall. As identities, like word in general, show a heavy-tailed distribution, we would expect worse performance from models that do not consider frequency information.

4.4.2 Case Study

Our case study briefly explores what can be learned by modeling both affective and semantic stereotypes of Twitter users. We first consider each form of stereotyping independently, and then explore how these two forms of stereotype help us to better understand one particular identity. All results below are given for model parameters at one model sample (900th Gibbs sample); results are nearly identical at all other samples considered.

Affective Stereotypes

Figure 4.2 displays the affective stereotypes learned by the model via the parameter vector μ. While any analysis of affective stereotypes is inherently subjective, model results fit at least our own intuitions. The most negatively viewed identities were rapists, cowards and murderers, the most positive were intellectuals, spouses and volunteers. The least powerful identities were hostages, addicts and prisoners, the most powerful were advocates, grads and entrepreneurs. The most active were gangsters, cheerleaders and little brothers, the least active were poets, Catholics and vegans. Additionally, we find a strong positive correlation between the evaluative and potency dimensions, a characteristic long found by Affect Control researchers (Rogers et al., 2013).

It is also interesting to compare affective stereotypes estimated by the model (μ) to the survey data we utilized as priors (μ_0). Most relevant to the Eric Garner and Michael Brown tragedies are identities relating to the police and to protesters. Our identities of interest included three "police identities" - cop, police officer and police - and one protester identity for which we had prior data - protester.[8] On average, police identities were viewed as being "more bad" (avg. difference between μ_0 and $\mu = -.39$), more powerful (1.07), while protesters were viewed as being "more good" (0.14) and less powerful (-0.32) by users in our dataset than what our survey data indicated. Police and protester identities were both viewed as being less active (avg. -0.78 and -.38, respectively). Results are consistent with the common narrative that portrayed police as the perpetrators of the tragedies and ensuing protests, with the protesters often cast as the victim. Though multiple vantage points existed on Twitter, this finding is consistent with the use

[7] the simple baselines improve 17.9% and 15.4% over random on the semantic and affective tasks, respectively. Further, while not shown, the affective baselines used here considerably out-performed models based on approaches derived directly from ACT as well as several other parameterizations of our model, suggesting that they are a useful basis for comparison.

[8] We also included the misspelled identity "protestor", however, priors for this identity were developed empirically

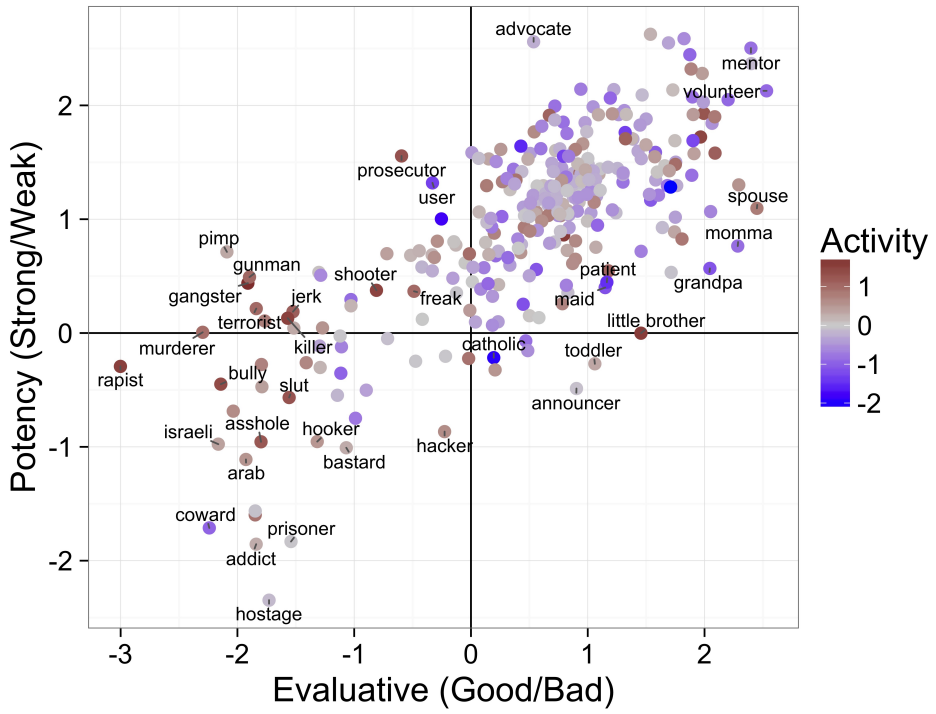


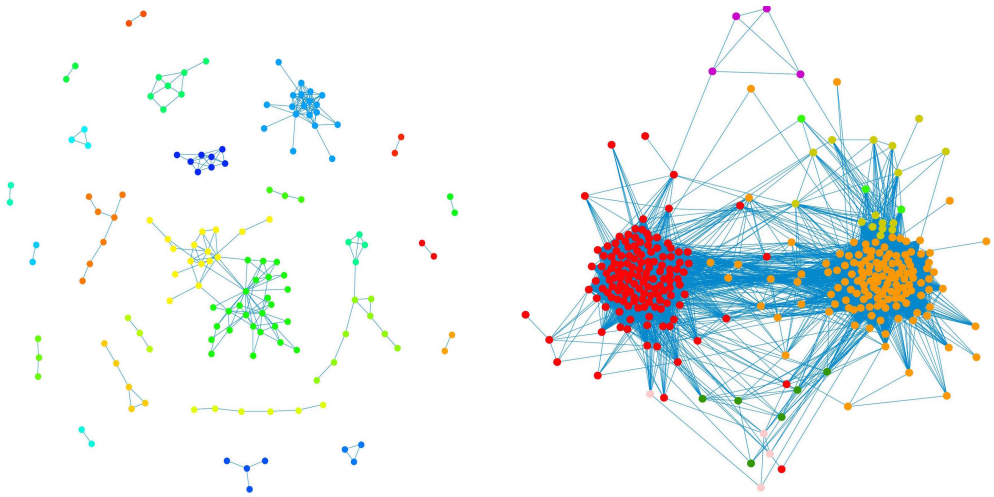
Figure 4.2: Affective Stereotypes of all 310 identities as measured by μ. The X- and Y- axes display the Evaluative and Potency dimensions, respectively. Color represents the Activity dimension. We only provide labels for outlier nodes in order to avoid clutter.

of Twitter to express frustrations with the police and sympathy for the protesters.

Semantic Stereotypes

The semantic associations of interest to us are captured in the parameter matrix Λ, the correlations in frequency of use between pairs of identities. We follow prior work (Blei and Lafferty, 2007) and use the Graphical LASSO (Banerjee et al., 2008; Friedman et al., 2008; Zhao et al., 2012) to sparsify Λ to better visualize and interpret sub-structures of the matrix. Figure 4.3 shows a network representation of Λ after applying two different levels of sparsification, as defined by the sparsification parameter λ. In each figure, nodes are identities of interest and are colored by their cluster in the network as determined by the Louvain method (Blondel et al., 2008). Links represent positive correlations between two identities.

Figure 4.3a) shows that when Λ is highly regularized (λ is high), strong, isolated institutional clusters of identities can be found. For example, network components exist in Figure 4.3a) representing grade school (freshman, sophomore and junior), the legal profession (attorney, judge, lawyer and prosecutor) and sports (coach, player, fan, athlete, announcer, qb, pitcher, teammate). Figure 4.3b) shows, however, that as sparsity decreases, identities form two main clusters. These two clusters can roughly be characterized by a split between identities we might use in an informal or social context (daughter, idiot, guy, friend) versus those we might use in a more news-oriented or formal discussion (e.g. republican, lawyer, priest). This split between the varying use of Twitter as both a platform for social interaction and information spread is well documented



(a) (b)

Figure 4.3: Network diagrams of semantic stereotypes as estimated by the model parameter Λ under various levels of sparsification, parameterized by λ. Sparsification levels of $\lambda = .6$ (a) and $\lambda = .3$ (b) are used. Isolates are removed from the image.

(Yang et al., 2012), as are the varying ways people use Twitter to present the self in both formal and informal contexts (Marwick and Boyd, 2011). It also suggests patterns in how people in different social settings find differential access to particular institutional settings (Heise, 2007).

Analysis at multiple levels of sparsification therefore shows hierarchical structure in semantic stereotypes consistent with prior work- identities are clustered into topics, or "institutions", that are themselves structured by how they use media and how they access institutions. These drivers of hierarchical structure differ, however, from the denotive constructs used to define hierarchies on Wordnet, suggesting that Twitter users may organize and interact with the social world around them in a decidedly "un-denotive" fashion.

Uncovering how these factors interact to produce the structure observed will require future work with models that allow multiple levels of network structure - to the best of our knowledge, such models blending hierarchical and network structure exist for social networks (Wang et al., 2013) but have not been applied to text. By modeling word correlations directly we were able, however, to observe this tiered network structure, suggesting the advantages of removing the assumption of strong topical clusters where little evidence exists to suggest the assumption is appropriate.

Stereotyping of "Thugs"

In addition to considering affective and semantic stereotypes independently, it is instructive to consider how their combination can help to interpret meanings of identities. One identity of interest to our dataset is thug. As John McWhorter notes in lieu of media coverage of the Freddie Gray protests in Baltimore, "...thug today is a nominally polite way of using the N-word...It is a sly way of saying there go those black people ruining things again".[9] McWhorter argues that

[9] http://www.npr.org/2015/04/30/403362626

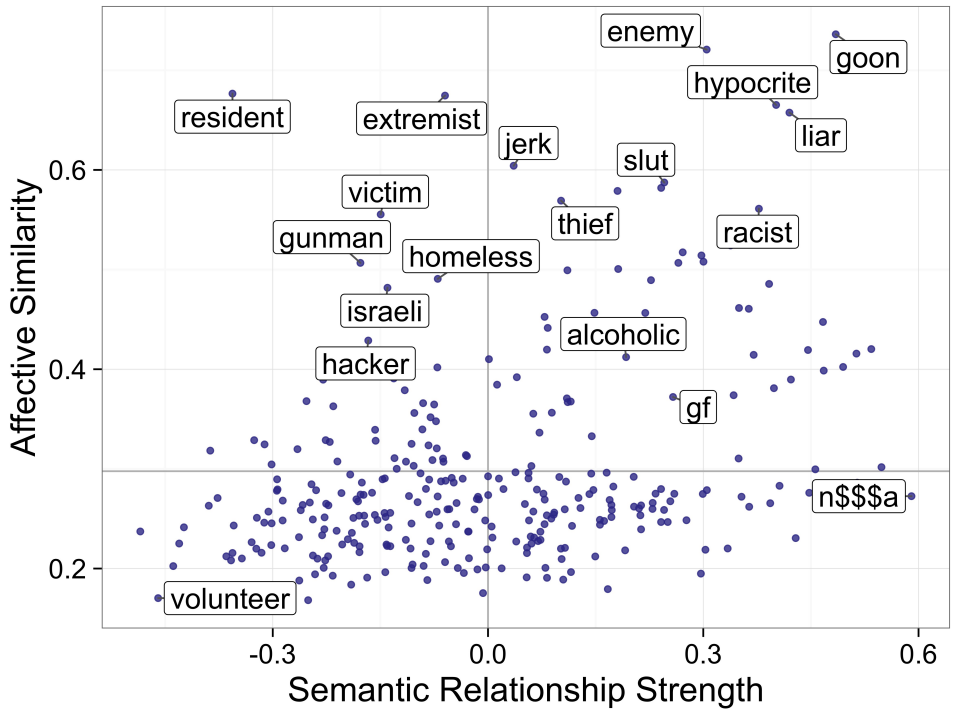


Figure 4.4: The x-axis gives each identity's measured semantic relationship to thug, the y-axis gives the identity's affective similarity to thus, computed as the unnormalized Euclidean distance between measured EPA profiles. Only outlier points are labeled. Grey lines on the x- and y-axes represent a null semantic relation and the mean affective similarity, respectively. A point is shown for all identities except for thug.

while the news media was using the identity thug to distinguish protesters who turned violent, the term has an underlying semantic association to the African American community.

If this semantic link is present in wider swaths of American culture, we should expect to see it emerge in social media data. Sure enough, the top five most similar words to thug that are not alternative spellings of the word (e.g. thugs) in the publicly available 200-dimensional GloVe word embeddings (Pennington et al., 2014) *trained on Twitter* are gangsta, ni$$a[10], goon, lil and homie. While McWhorter's argument focuses on a different "n-word" and makes a case for semantic similarity rather than association, his viewpoint nonetheless is relevant. There is a clear blend of affectively negative identities (goon) intermingled with identities that either refer to or are often used by the African American community.

Our model also retains the connection between thug and identities related to the African American community. However, as Figure 4.4 shows, by separating out semantic stereotypes from affective stereotypes, our model allows one to more readily discern nuanced relationships between thug and other identities. Our model agrees with the GloVe embeddings in that thugs are "culturally synonymous" with goons - that is, these two identities have similar affective meanings and strong semantic similarity.

However, we see that thugs are semantically related but not affectively similar to the identity

[10]we have replaced the letter "g" with $

"ni$$a", suggesting that while similar people refer to these two identities frequently, it is done so in different emotional contexts. Affective meaning may in this way be a useful tool to help differentiate between similarity and relatedness - words people feel differently about are unlikely to refer to the same or similar concepts. Figure 4.4 also suggests that uncovering affectively similar but semantically distinct identities may provide a way to "map" institutional settings onto one another. For example, thugs and extremists are not semantically related but are affectively so - perhaps thugs are the extremists of their own particular institution.

4.5 Conclusion

We present a new Bayesian model of text grounded in the socio-cognitive theory of stereotyping. To the best of our knowledge, it is the first large-scale exploration of stereotyping as defined by both the way people feel about identities as well as how they (implicitly) suggest they are semantically associated. Validation experiments show our model is useful for predicting the identities people will use in a given tweet. We used the model to explore both semantic and affective stereotypes of identities in a dataset of 45K Twitter users who actively discussed the Eric Garner and Michael Brown tragedies.

Users in our dataset viewed police-based identities more negatively and the protester identity more positively than prior survey-based data would have led us to believe. This fits with one of the stronger narratives of the Eric Garner and Michael Brown tragedies. On a broader scope, we found that semantic stereotypes revealed a hierarchy of identity topics, or "institutions", that fit into well-established notions of institutional settings in the sociological literature. These institutions themselves were clustered, revealing a hierarchical structure of identities distinct from traditional identity hierarchies, such as those found on Wordnet.

Finally, we observed that differentiating between affective and semantic stereotypes allowed us a more nuanced view of stereotypes of the identity thug. This latter point suggested that it may be easy for (what we consider to be) undesireable stereotypes to "creep in" to computational models of word meaning. One remedy for this may be to explicitly differentiate affective meaning from semantic meaning, as recent work has begun to do (Maas et al., 2011; Tang et al., 2014). Similar care must be taken in interpreting the semantic hierarchies of identities uncovered by our method and those it builds on. For example, one cluster in Figure 4.3c) represents the three identities "Muslim", "extremist", and "terrorist". While this cluster may accurately reflect unfortunate semantic connections between these identities in contemporary American culture, we would claim this to be an undesirable "belief" of contemporary AI systems. How to train systems to avoid these two kinds of undesirable biases should be a focus of future work; incorporating affective meanings (Maas et al., 2011; Tang et al., 2014) and blending rule-based systems with neural embedding models seem like promising starts (Liu et al., 2015).

The Twitter data that we use is biased in important ways (Malik et al., 2015; Ruths and Pfeffer, 2014), and thus our findings should not be seen as generalizing even to the entirety of Twitter, let alone to more general social settings. Further, as results for semantic stereotypes suggested, using Twitter likely gives us a mix of explicit and implicit biases (Greenwald and Banaji, 1995), as we measure a blend of a users' normative environment when they tweet as well as their underlying views. However, results from our case study suggest that thinking about identities in semantic *and* affective ways, can help us to characterize role relationships, to better

understand differences between explicit and implicit biases, and to see where methods naïve to this distinction may introduce prejudices into word embeddings.

These dangers are important because identities are special - they convey affective and semantic stereotypes that have served as the basis for behaviors from bullying to genocide. Miller and Charles (1991) seem, at least on some level, to agree with us that identities are important, stating that "[p]resumably, person terms are treated differently because language is spoken by people, to people, and (often) about people". We look forward future work combining NLP and socio-cognitive theory to confirm or deny the implications we have put forth and to improve upon the methods presented.

Chapter 5: Exploring How we Label Other People

5.1 Introduction

Identity labels are the words or phrases we use to describe our social roles, categories and group memberships (Smith-Lovin, 2007; Tajfel and Turner, 1979). While not all roles, categories or group memberships can be associated with a label (see, e.g. discussions in Penner and Saperstein, 2013), these discrete linguistic categories are widely used by humans to inform our actions (Heise, 1987). Across the social sciences, it is generally accepted that the way we label ourselves, the way we are labeled by others and the way we label those around us all impact our behavior.

What is less clear, however, is how we choose the labels to apply. In other words, given a particular individual in a particular situation, an active area of study is in the development of models that can predict how she will be labeled, either by herself or someone else. Social psychologists have studied both how we label ourselves (Burke, 1980; Davenport, 2016; Miles, 2014; Owens et al., 2010; Stryker and Burke, 2000) and how we label others (Heise, 1987, 2007; Johnson et al., 2012; Kang and Bodenhausen, 2015; Penner and Saperstein, 2013; Robinson et al., 2006; Schröder et al., 2017). Neuroscientists (Cikara and Van Bavel, 2014; Van Bavel and Cunningham, 2010), cognitive psychologists (Freeman and Ambady, 2011; Kunda and Thagard, 1996; Read and Miller, 1998; Schröder and Thagard, 2014), linguists (Bucholtz and Hall, 2005; Recasens et al., 2011) and even computer scientists (Yao et al., 2013) have also focused on variants of this identity labeling process.

Perhaps the most developed quantitative sociological model that can be used for identity labeling is *Affect Control Theory* (ACT) (Heise, 1979, 2007). ACT postulates that identities are defined by a set of affective attributes and that these attributes, in connection with a cybernetic control system, are used to make sense of social situations. While ACT is not necessarily meant to explain how identities are selected (Heise and MacKinnon, 2010, pg. 200), its mathematical model can and has been used to understand how identities engaging in social events will be labeled (Heise, 2007). For example, the mathematics of ACT can be used to predict that teachers are more likely to "instruct" interns than thieves, and that mothers are more likely to hug their children than their enemies.

Where ACT falters as an identity labeling model is in its ability to address the strong *semantic* relationships that exist between identities. Intuitively, we expect *teachers* to instruct *students*, not *interns* as the theory predicts. ACT's exclusion of these "semantic intuitions" restricts the theory to predictions that make sense, but that are not "correct". Heise and his colleagues(Heise,

1987, 2007; Heise and MacKinnon, 2010) have explained these semantic relationships and the restrictions they place on identity labeling through the concept of *institutions*, which "organize the huge number of identities that [one] can encounter" (Heise, 2007, pg. 28). However, such semantic relationships have not been formally introduced into ACT's mathematical model. Despite the pervasive effects they are expected to have on identity labeling, institutions therefore remain a proscriptive, blurry factor in ACT.

This issue is further compounded by the tenuous assumption that information on semantic relationships is retained only within institutional structures. As Lizardo and Strand (2010) argue, it is instead more plausible that these semantic meanings of identities, and cultural forms more broadly, are also lodged within our cognition. To this end, semantic relational models of identity labeling have been developed by cognitive psychologists. In particular, recent connectionist[1] interpretations of the identity labeling process explicitly introduce cognitive mechanisms that explain how semantic relationships enable and constrain the identity labeling process (Freeman and Ambady, 2011; Kunda and Thagard, 1996). These cognitive models have been shown to be strongly predictive of identity labeling processes, thus making them in some senses more "correct" than ACT.

However, these cognitive models are also problematic. Incorporating affect into existing connectionist models must either be done by assuming affective profiles implicitly exist for each identity or by making explicit connections between emotional profiles of identities, which reduces the interpretability of such models (Schröder et al., 2014). More importantly, parameterizations of these models apply to particular social situations (e.g. particular experiments) and are hand-crafted for each theoretical task. These cognitive models thus lack both the parsimony and wide-ranging applicability of Affect Control Theory.

The task of this chapter is to explore an integration of these identity labeling models. We will do so in several steps. First, we develop a generic framework for expressing the identity labeling problem mathematically in a way that encapsulates other similar models as well. Second, we develop a basic, parsimonious instantiation of this generic framework that expresses three key concepts of *relationships between identities* that have been shown to be important in the identity labeling process in prior work. The first concept is *semantic similarity*, which we define as the extent to which two identities can be used to define the same person. The second concept is *semantic association*, the extent to which two identities are seen in the same social context. The third is *affective similarity*, the extent to which we feel the same way about two identities.

In order to test whether or not these factors actually play a role in identity labeling, I first develop measures for each of these concepts from existing, publicly available survey data for 87 different pairs of identities. I then develop and instrument a survey in which respondents perform identity labeling by providing answers to multiple choice questions in two distinct, text-based, hypothetical social situations, examples of which are given below:

- Who would you say is most likely to be *seen with* a teacher?
- Given that someone is a teacher, what other identity is *that same person* most likely to also be?

The two questions above are similar but are expected (and shown) to have different mechanisms by which individuals determine the appropriate label. In the first question above, for

[1] More specifically, *parallel constraint satisfaction* models, to be detailed below

example, people would be likely to respond with "student". However, this is among the least likely answers to be given in the second question, as "teacher" and "student" are mutually exclusive role identities. While these questions shrink the complex process by which identity labeling occurs down to a series of survey questions, they therefore are useful as a starting point for understanding the importance of these three factors in the identity labeling process.

We observe in this first survey that semantic factors are very important in determining the outcome of the identity labeling process. In order to confirm these results, I also develop a second survey for which measures of semantic similarity are not available but for which the selection of identities is less biased. The surveys provide complementary sets of information that are useful in better understanding how these factors impact identity labeling. However, what the survey results also suggested was that the existing measurements of semantic similarity and semantic association were insufficient for a variety of reasons. Most importantly, the data we used could not tell us *how* identities were semantically similar or semantically association.

To this end, I propose a mathematical, theoretical model that defines how we should think about measuring these questions. The model is inspired from the sociological side by ACT and Burke's (1980) seminal work on measuring identity and from the cognitive side by the connectionist approaches of (Freeman and Ambady, 2011) and the seminal work of Tversky and Gati (1978). It provides a new representation of stereotypes that unifies these different views of what a stereotype is and how to measure them. I provide a way to estimate the model using the second round of survey data I collected and explore initial results from that work.

5.2 Related Work

The study of identity labeling spans at least the disciplines of sociology, social psychology and cognitive psychology. As we cannot cover all relevant material across these fields, we focus in this section on clarifying where our model and methods fit in to and draw from the existing literature across these domains.

5.2.1 Sociological and Social Psychological Models of Identity

Quantitative sociological models of identity are generally interested in characterizing the attributes of particular identity labels in terms of traits (Burke, 1980) or latent affective spaces (Fiske et al., 2002; Heise, 2007), and how these attributes of identities influence human behaviors (Cuddy et al., 2007; Heise, 2007). On some level, an implicit assumption of such work is that stereotypes, "overgeneralizations about a group or its members that are factually incorrect and inordinately rigid" (Dovidio and Gaertner, 1999, pg. 101), and prejudices, "an unfair negative attitude toward a social group or a member of that group" (Dovidio and Gaertner, 1999, pg. 101), exist either on a unidimensional scale of good/bad or can be expressed as a representation of a particular set of traits. Several attempts have been made, however, to formalize or extend these assumptions into a mathematical model of how these perceptions are internalized.

The Stereotype Content Model (SCM) (Fiske et al., 2002) theorizes that stereotypes of identity labels exist within a two-dimensional affective space. The axes of this space define the "warmth" and "competence" of a particular identity. The SCM tends to focus on perceptions of social groups as opposed to focusing on role identities. The warmth an individual feels towards a particular group is presumed to be a function of the extent to which the group competes with the

individuals own group. The competence an actor perceives of another group is a function of the other groups level of status relative to the individuals own group. Fiske et al.'s (2002) original statement of the SCM utilized survey data to explore these perceptions of a variety of identities, and Cuddy et al. (2007) has provided a theoretical model of how these perceptions may drive emotions and action.

In the SCM, the affective meaning of identities along the warmth and competence dimensions are assumed to be universal to the extent that, when asked "what do you think people in general think about identity *X*", individuals within a particular "national culture"[2] will give similar responses. The SCM predicts that specific stereotypes will lead to specific prejudices, defined as emotionally biased opinions towards a member of a social group. High warmth and high levels of competence lead to admiration, low warmth and low competence lead to contempt, low warmth and high competence lead to envy and high warmth and low competence lead to pity. Finally, Cuddy et al. (2007) posits that these prejudices will predict behaviors oriented towards members of groups.

Recently, Rogers et al. (2013) compared the attributional space presumed by the SCM to the dimensions of affective meaning along which identities are measured in Affect Control Theory (ACT) (Heise, 2007), a sociological model of identity and action. In ACT, each identity is defined by a three-dimensional affective profile. The first dimension, the Evaluative dimension, specifies how "good" or "bad" an identity is. The Potency dimension specifies the powerfulness/powerlessness of the identity, and the Activity dimension defines the activeness/passiveness of the identity. Combined, measurements along these three dimensions for a particular identity are referred to as that identity's EPA profile. ACT, like the SCM, assumes that the EPA profile of an identity are generally universal, though stark differences indicating particular subgroups of society do exist Smith-Lovin and Douglas (1992).

Rogers et al. (2013) finds that the warmth and competence dimensions of the SCM show strong correlations, respectively, with the Evaluative and Potency dimensions of ACT. For the purposes of the present work, then, it is useful simply to note that the idea that identities can be represented by their affective meanings has received significant support. With respect to ACT, these profiles have been used for a variety of purposes, for example, predicting how individuals will behave in particular situations Smith-Lovin and Douglas (1992). However, recent years have seen significant advancements to the existing underlying mathematical model of the theory, thus altering some of its substantive claims.

First, recent work has focused on addressing the fact that ACT in its original form focuses only on point estimates of EPA profiles. New models have been developed with the assumption that variance around these estimates may exist, and that this variance may help us to understand uncertainty in social situations (Schröder et al., 2017). This observation has led to the development of *BayesACT*, a probabilistic interpretation of the original ACT model (Hoey et al., 2013b). Others have criticized ACT for lacking a strong theory of the self. In response, Heise and MacKinnon (2010) developed an Affect Control Theory of the self, which was subsequently incorporated into a Bayesian theory of the self by Hoey and Schröder (2015).

In the adoption of a strong theory of the self, ACT draws comparisons to other prominent theories of the self, including social identity theory Tajfel and Turner (1979) and its predecessor

[2]For the moment, we sidestep a discussion of whether or not such a thing exists

self-categorization theory Turner et al. (1987), as well as to identity theory Stryker and Burke (2000) and its predecessor, identity control theory Burke (1980). Owens et al. (2010) provides a valuable overview of these theories and how they help to predict how an individual will identify herself in a social situation.

As the present work focuses on how we label *others* as opposed to the self, these theories are not directly applicable. Further, only ICT provides a quantitative model of identity meaning. ICT is useful in this way, as it presents a well-known argument for the measurement of identity meanings with an explicit focus on identity relations. ICT's measurement model of an identity label is based on the set of dynamic characteristics representative of the self. These characteristics, which differ from study to study, are not traits (which are considered static), but rather dynamic aspects of an identity which are variously shown in different social situations (Stets and Carter, 2012). In this way, ICT's measurement model emphasizes the fluidity of identity.

The measurement model of ICT is also developed with the assumption that identities are defined, in part, by their relations to other identities. ICT's measurement model thus results in "a semantic space" where "[i]dentities located close together [have] similar action implications, while identities located more distantly from each other would have very different action implications" (Burke, 1980, pg. 22). The mathematical model we develop is based on similar notions of a latent space, much like ACT and a significant amount of work in other veins of the mathematical sociology literature, perhaps most notably that of Blau (1977).

ICT was developed in large part from Burke's (1980) perspective of how identities should be measured. Our model adheres closely to the conceptual ideas presented in this work. Most notably, we adhere to the idea that identities are relational. On this point, Burke (1980) discusses the idea of *role/counter-role* pairs, i.e. those between "brother" and "sister", and generalizes this to the idea of *identities and counter-identities*, the notion that identities in general have one or more other identities on which we base their meaning. In our model, these counter-identity pairs will serve as bases for the dimensions along which measurement occurs.

In ACT, these kinds of semantic relationships between identities are defined via the conceptualization of institutions, which are essentially relations in the form of clusters of identities that can be grouped together via institutional settings. Heise and MacKinnon (2010) use network text analysis to show that identity labels cluster into "institutions", and use this concept of institutions to describe how certain identities are implicitly more probable in particular settings.

However, ACT's definition of the institution and how individuals determine institutions from contextual cues are intentionally ambiguous. Institutions and their link to identities therefore do not exist within either ACT or BayesACT's mathematical model. It is for this reason that artificial agents in simulations using the BayesACT model can revert to situations like the following, quoted from Schröder et al. (2017): "...both agents have developed the shared belief that one of them (agent A) is an "executioner" while the other (agent B) is a "great grandmother". While *affectively*, this may have made sense given initial constraints, from a relational perspective, this identity pairing is significantly unlikely.

In ignoring the ability of the human mind to internalize institutional boundaries, ACT also ascribes too much power to institutions in defining appropriate identities for a particular situation. Taking from Lizardo and Strand (2010) *strong practice theory* the assumption that what is internalized within institutions is also largely internalized within the human mind, we adopt a cognitive approach to modeling relationships between identities. Through the vehicle of cogni-

tive modeling, the present work incorporates this relational knowledge assumed away to institutions by Heise and MacKinnon (2010) and the affective attributional meanings of ACT. A similar cognitive argument for the principles of ACT is made by Schröder et al. (2014), who note that the three dimensional representation of identity labels made by ACT is an appropriate representation for attributes of a *schema* of identity labels.

5.2.2 Cognitive Models of Identity Meaning

A long line of research exists related to the study of stereotype and prejudice in cognitive psychology - for an older but high quality review, we point the reader to Greenwald and Banaji (1995). Our focus here is on recent models of identity labeling and stereotype in the cognitive psychology literature that are relational in nature. By this, we mean that in these models, parameterizations are defined by where *semantic links*, or generic cognitive associations, do (or do not) exist between identity labels.

In the works referenced here, semantic links are constructed as the basis for *parallel constraints satisfaction models* (here, PCS models for short). Semantic links, which are the named "constraints" in PCS models, exist between nodes, which themselves can be thought of as cognitive schemas. In PCS models, each node is ascribed a single attribute - a level of activation. Each node in the model starts with a base level of this activation value. Nodes can then be externally "excited", at which point their base level of activation is increased by a set amount. At this point, activation then flows through links in the system. These links can either act as exciting links or inhibitory links. An exciting link between two schema means that activating one of the links will increase the activation of the other. An inhibitory link means that increasing activation in one schema will *decrease* the activation of another. The existence of these inhibitory links is the primary benefit of PCS models over pure spreading activation models (Collins and Loftus, 1975).

Several PCS models have been developed that help to understand how semantic relationships inform the identity labeling process (Freeman and Ambady, 2011; Kunda and Thagard, 1996; Schröder et al., 2013), we review two in detail here. Schröder et al. (2013), focusing on a model of behavioral priming, develop a PCS model that draws directly on ACT. In their model, the nodes of the system are Evaluative, Potency and Activity dimensions for different identity labels. Links are drawn based on empirical data and the model is used to simulate various experiments in which individuals were cued with stereotypical information that mediated their behaviors in social situations. In many ways, Schröder et al. (2013) model achieves the combination of sociological and cognitive models that we aspire to here. However, we prefer the model presented in the present work because treating EPA profiles as attributes of nodes, rather than as the nodes themselves in a PCS model allows us to retain discussions of identity relationships (e.g. role relationships) at the same level as our modeling framework. Because of this, we are able to leverage existing datasets to parameterize our model, rather than relying on hand-crafted parameterizations that validate particular experiments.

The work of Freeman and Ambady (2011) provides a model closer to the ideals discussed in the present work. Their model seeks largely to link research on person construal, which "seeks to examine the lower-level perceptual mechanisms and determinants of categorization", with models of more abstracted cognitive processes, such as prejudice. In their model, nodes can be one of four types. Nodes at the cue level include visual and auditory features, such as an

individual's face. At the category level, nodes indicating social categories (e.g. gender) exist. At the stereotype level, nodes exist that represent traits, such as annoying. Finally, a "higher-order" level includes nodes such as prejudice and motivations. Connections exist across nodes at different levels, and activation starting anywhere in the network is passed through the network until stability is reached, at which point an identity (category) is probabilistically selected based on its level of activation.

Kunda and Thagard's (1996) model, a predecessor of Freeman and Ambady (2011), is also a PCS model. Kunda and Thagard (1996) model traits and stereotypes, attempting to discern how people make decisions about actions based on information presented to them. In both Kunda and Thagard's (1996) and Freeman and Ambady's (2011) work, several advantages of PCS models are clear. First and foremost, an explicitly relational perspective of identity makes it much easier to model how overlaps and intersections between identities, both implicit and explicit, may manifest during the labeling process (Penner and Saperstein, 2013). While PCS modelers generally do not take a stance on how semantic relationships develop over time Schröder et al. (2014), cognitive associations can be used to explain why, for example, Asian individuals are more likely to be assumed female than individuals of other races . Further, varying the strength and valence of these interrelationships can help to explain how, for example, biracial women can be more likely to identify as multiracial than biracial men (Davenport, 2016).

The use of a cognitive, relational perspective therefore allows PCS models the ability to provide an appealing explanation for a variety of well-known phenomenon in the identity labeling process, in particular the existence of intersectional identities. However, there are three important limitations of PCS models. First, parameterization (i.e. the definition of links between identities) is done largely by surveying prior literature and via the researcher's use of intuition. While such methods are appropriate for smaller scale studies of identity, attempts to set positive and negative links across a larger scope of identities, like the set studied in the present work, quickly becomes unweildly.

Second, modeling affective information, how individuals *feel* about identities, is difficult. PCS models rely on the implicit notion of an identity's affective meaning (e.g. implicitly, criminals are bad), but lack the expressiveness to represent this explicitly in the model. Given the importance of affect and emotion in our labeling of others demonstrated by ACT, a more appropriate model would thus retain some element of affective meaning of identities.

Finally, PCS models are *context-specific*; that is, they are designed to be hard-wired to how people label others in particular situations. This is because only one type of semantic relationship can be modeled in a given PCS model. A single model therefore cannot be used to demonstrate how the same individual, when asked two different questions about how to label a particular individual, may use the fact that a role relationship exists between a brother and a sister in particular ways. That is, PCS models cannot represent the fact that two distinct individuals standing together may be brother and sister (thus an "exciting link" exists between these two identities) and, simultaneously, that a single individual is very unlikely to be *both* a brother and a sister (thus an "inhibiting" link should exist).

Fortunately, other psychologists have provided a framework for how to think about these different kinds of semantic relationships (Collins and Loftus, 1975; Tversky and Gati, 1978). These ideas have been adopted heavily in recent years by computational linguists seeking to extract semantic meaning from text (e.g. Resnik, 1999). In the following section we review the

two most important types of semantic relationships of interest to the present work - semantic association and semantic similarity.

5.2.3 Measuring Semantic Relationships

Computational linguists often find it important to distinguish between semantic *association*[3] and semantic *similarity* Hill et al. (2014a). Semantic associations are generic cognitive connections we make between concepts; i.e. those connections that are relevant in free association tasks (Nelson et al., 2004) and that are generally represented in spreading activation models of cognition (Anderson, 1983, 2007; Collins and Loftus, 1975). As such, semantic associations tend to form between things we see together in particular situations - tables and chairs, for example. In contrast, semantically similar entities are those that refer to entities have have some common inherent property, perhaps best ascribed first to Tversky and Gati (1978). Semantic similarities arise when concepts have similar properties - for example, we can drink coffee out of both a mug and a cup. Resnik (1999) differentiates between similarity and association via the following example: "cars and gasoline [are] more closely [associated] than, say, cars and bicycles, but the latter pair are certainly more similar".

The analogy used by Resnik (1999) can be modified to partially address the issue relating to PCS models described above. In the case of labeling a person standing next to a brother, we are likely to rely on semantic association - we tend to label people as brothers in situations where sisters are also present, this eventually produces a strong association between these concepts. In contrast, "brothers" have the distinct property, as being viewed as men, whereas "sisters" are typically women. Thus, to label the same entity as being both brother and sister would disagree with an important notion of similarity as we view people.

The distinction between similarity and association is roughly analogous to Burke's (1980) work in that "counteridentities" and "counter-roles" can be seen as highly associated, while those with similar dynamic characteristics can be seen as highly similar. The difficulty in each case, however, lies in defining exactly what properties are important in defining similarity. This is problematic even in the example presented; specifically, while brothers and sisters are rarely the same person because they have different *gendered meanings*, they both have the important property of being types of *siblings*. We will develop in this chapter a mathematical model that addresses this distinction by assigning variable weights to these different types of properties that identities may hold - in this case, gender is a stronger indicator of dissimilarity than sibling is of similarity.

At this point, it is necessary to clarify further why we are interested in connecting our work to the efforts of computational linguists. The primary reasons for this connection is that over the past three decades, computational linguists have focused heavily on *measuring* semantic similarity and semantic association Agirre et al. (2009); Charles (2000); Hill et al. (2014a); Levy and Goldberg (2014); Mikolov et al. (2013a); Miller and Charles (1991); ZHANG et al. (2013). These measurement models, and how they are validated, are useful in helping to understand the current scope of meaning of similarity and association in NLP and how work specific to identity must move beyond these definitions.

Computational linguists use two high level approaches to measuring similarity and associa-

[3] also called semantic relatedness (Resnik, 1999)

tion. The first approach, called the *distributional*, or *corpus-based* approach, uses large datasets of text and a vast array of methods to compute the extent to which words are associated or similar. The second, the *knowledge-based* approach, uses existing lexical databases to construct estimates of relatedness and similarity. We review both approaches here.

Corpus-based approaches to measuring similarity and relatedness

Early researchers seeking to use corpus-based approaches to extract semantic associations developed Latent Semantic Analysis (LSA) (Deerwester et al., 1990), which treats each document in a corpus as a "bag of words". LSA uses singular value decomposition (SVD) to characterize each word in a corpus in a low-dimensional space. In this space, words that *co-occur*, or appear together, in many documents will be close together. LSA has recently been supplanted by Bayesian admixture models of text, notably latent Dirichlet allocation (Blei et al., 2003) (LDA), which represent words as a mixture over latent "topics". In LDA, words that frequently co-occur in the same documents will be more likely to be in the same topic.[4]

Such methods have become popular in both the computational linguistics literature, where thousands of approaches have build on top of LDA (e.g. Blei and Lafferty, 2006, 2007; Li and McCallum, 2006; McCallum et al., 2007; Teh et al., 2012), and in the social science literature, where scholars have used the model to, e.g., extract themes from literary corpora.[5]. Such models have also been connected to the study of identity - Joseph et al. (2016b) observes that topic models can be used to find the latent institutions of identities found originally using network text analysis by Heise and MacKinnon (2010).

LDA, LSA and other related models using the document as a "bag of words" extract semantic associations via *syntagmatic contextual* relationships Sahlgren (2006) between words in a text, relationships built explicitly on co-occurence. In contrast, one can also consider *paradigmatic* relationships between words, that is, the extent to which one word can be "substituted" in for another words within, say, a sentence Sahlgren (2006). Models built on this paradigmatic perspective have, at least in principle, the ability to measure the semantic similarity of two concepts. Such models have drawn significant attention in recent years with the advent of "deep learning", where researchers train, e.g., recurrent neural network models on sliding windows of large text corpora (Mikolov et al., 2013a). These models general *neural embeddings* of words, which represent each word in a given vocabulary as a vector of continuously-valued numbers. These vectors can be compared between any two words to get a rough estimate of the semantic similarity between the two words.

Knowledge-based approaches to measuring similarity and relatedness

The *knowledge-based* approach to measuring similarity and relatedness leverages large, manually curated datasets of concepts and their relations. Wordnet (Miller, 1995) is a large dataset that contains lexical relationships of tens of thousands of concepts in the English language. Importantly, Wordnet represents concepts, called *synsets*, rather than words (or *surface forms*. Each concept may have multiple *lemmas*, which represent surface forms used to express the concept. Thus, each *synset* may have multiple *lemmas* - the synset for the concept of adult male is associ-

[4] A better but more jargon-y characterization is that the words will have similar distributions over the set of latent topics.

[5] See the special issue of Poetics on topic modeling Mohr and Bogdanov (2013)

Parameter	Interpretation
I	universe of all possible identity labels
S	The set of all institutional settings being modeled
P	The set of all dynamic characteristics, static traits and external cues being modeled
$\mathbf{i_y}$	The set of identities applied to a person y
p_x	Probability model conditioned on an individual x's current perceptions
c	*Observable* features of a context in which an identity labeling occurs - for example, the institutional setting
$\phi(i, y, c)$	A function used to determine how likely a particular identity i is for particular individual y in a particular context c

Table 5.1: An overview of the variables introduced in this section

ated with the lemmas "man" and "adult male". Similarly, the word "bat" is a lemma of multiple *synsets*, including one relating to the object used to hit baseballs and the animal that sleeps in caves. Synsets in Wordnet are connected by a concept hierarchy. A concept that is above another concept in this hierarchy is the *hypernym* of that concept, the opposite relationship is called a *hyponym* relationship. Thus, for example, e.g. professional is a hypernym of doctor, and doctor is a hyponym of professional. Other lexical relationships, including synonymy, antonymy and words that are "similar to" each other are also encoded in Wordnet.

Wordnet is the most commonly used lexical database for measuring semantic relatedness and semantic similarity, although studies tend to focus more on similarity- see Agirre et al. (2009) for an overview of some recent approaches. Models built on Wordnet generally perform better on prediction tasks on "gold-standard" datasets - that is, they are better able to match human judgements of semantic similarity than distributional approaches (Hill et al., 2014a; Le and Fokkens, 2016). However, some success has also been had by combining knowledge-based and corpus-based methods for measuring semantic relatedness and semantic similarity (Resnik, 1999) - some of these models are utilized in the present work.

In addition to lexical databases like Wordnet, researchers have also begun to utilize *knowledge bases* (Medelyan et al., 2013), large data source providing (semi-)structured information on millions of real-world concepts, to measure semantic relationships between concepts (e.g. Gabrilovich and Markovitch, 2007). The canonical example of a knowledge base is Wikipedia, but other examples include Yago2 (Hoffart et al., 2011, 2013), NELL (Mitchell et al., 2015) and Wikidata (Vrandei and Krtzsch, 2014). While lexical databases include high-level relationships between concepts like synonymy, knowledge bases are able to capture more complex, contextual relationships. For example, an "is-a" relationship in Wikidata between man and brother and sibling and brother, and an "opposite of" relationship exists between "brother" and "sister".

5.3 A Base Mathematical Formulation of Identity Labeling

In this section, we introduce a base mathematical formalization of the process of identity labeling. In Section 5.7, I will introduce an instantiation of this base model that encapsulates several existing models. The variables used are listed in Table 5.1 with a brief description of each. To

begin, we formally state the problem we are interested in quantifying:

> **The identity labeling problem:** Given a "labeler", person x, and a "labelee", person y, in a particular context c, determine which identities, $\mathbf{i_y}$, in the universe of all identities I that x will apply to y

Here, "context" could be any information available to the labeler that is not given, implicitly or explicitly, by the labelee. The context thus might include information about other people in the current setting or information about the setting itself (Smith-Lovin, 1987a). The statement of this problem thus inherently assumes that the only information relevant to x when deciding on an appropriate set of labels for y is other information x knows about y or information that can be construed as context in the present situation.

We further assume an element of randomness is inherent in how we label others- even in two very similar contexts and with similar information about a person, we might for some reason choose two different labelings. Because of this randomness, identity labeling is best described as a probability distribution, where each identity has some (perhaps very small) chance of being applied at any time. Formally, we build a model for the probability that the labeler x labels y with the set of identities $\mathbf{i_y}$ given some information on y (say, her age) and information about the context c (say, that the setting is a school yard):

$$p_x(\mathbf{i_y}|y, c) \tag{5.1}$$

This expression says that every person x will have a different way of using information about y and c in coming up with a set of labels that they think is appropriate. However, in the present work we will assume that $p_x = p \,\forall\, x$; in other words that all people have the same perceptions of the world and thus will label person y in the same way given the same information. This assumption is, of course, hopelessly incorrect, but it is made in the present work for two reasons. First, the survey data that we collect is not large enough to estimate per-individual statistical models. Second, both Affect Control theorists and cognitive psychologists generally make similar assumptions for similar reasons.

In the present work, we will also assume that $\mathbf{i_y}$ is always of size one - that is, that x will provide y with exactly one identity. This assumption makes it possible to develop single-answer multiple choice questions in a survey, lessing the cognitive demand on our respondents. Finally, we will also assume that we know the set of all identities that x could possibly apply to y, given by the set I as described in Table 5.1. These assumptions allow us to re-express the probability model in Equation5.1 as the following *discrete choice model* (McFadden, 1980):

$$p(i_y|y, c) = \frac{e^{\phi(i_y,y,c)}}{\sum_{j\in I} e^{\phi(i_j,y,c)}} \tag{5.2}$$

In this expression, ϕ is any function that returns a value based on a particular identity label (e.g. i_y or i_j) and information about y and c. Equation 5.2 therefore states simply is that the odds that x chooses a particular label for y are relative to the odds that they select any other possible identity in the set I, which contains all possible identity labels. We will develop a specific instantiation of this general model by assuming a specific parametric form of $\phi i, y, c$. To do so, we will use the additional assumption that the variables y and c are expressed as identity labels, i_y and i_c. That is, where information about y is given, it is given in the form of another

label that person y already has (i.e. a "student"). Where information on context is given, it is given as the identity, i_c, of another individual (that is not y) in the given context.

This assumption is not strictly necessary- Equation 5.2 can be generalized to much broader classes of assumptions. For example, Hoey et al.'s (2013b) BayesACT is expressed as a restricted form of this model. However, using this assumption we are able to incorporate prior work on both identity and semantic relationships between words in general that are described or can be cast in a relational manner - ACT's institutions and affective profiles, Burke's counter-identities and dynamic characteristics and contextual cues, Freeman's static traits, semantic association and semantic similarity can all be cast as a relationship between two identities i_a and i_b.

Further, many of these factors can be subsumed within one another. The idea of semantic associations incorporates the concept of the counter-identity as well as shared institutions. Semantic similarity encapsulates both the dynamic characteristics used by Burke (1980) and the static traits used in Freeman and Ambady's (2011) model. Affective similarity could also be considered to fall under the umbrella of semantic similarity- however, in the present work, we choose to keep it separate in order to explore its effects in ways discussed below.

We therefore assumes three distinct classes of factors impact the identity labeling process - semantic similarities between identities, semantic associations between identities, and affective similarities between identities. Semantic associations should be important in determining which identities will co-occur in the same social situation. For example, if told that another individual in a particular context is known to be a soldier, the identity of the person standing next to them is likely to be some kind of military identity as well. In contrast, semantic similarity should be important in determining whether or not two identities are likely to be applied to the same individual. Finally, affective similarity may be used in determining identities in both of these cases.

In both study one and study two, we consider two different kinds of (hypothetical) social situations. Importantly, we will *assume that the attributes and relationships between identities do not change, it is only how this information is used that changes between these two social situations*. Thus, for example, the semantic similarity between identities is a consistent value across all social situations, but semantic similarity is used differently when determining multiple identities of a single individual versus pairs of identities that can be expected to be together. Consequently, we can develop a conceptual model applicable to many social situations by thinking only about the importance of different facets of an identity's underlying meanings rather than attempting to reconstruct new meanings for each question type.

In the present work, this means thinking about how semantic similarity, semantic association and affective similarity impact the two different types of survey questions we present. Let us operationalize ϕ as a pairwise function, $\phi(i_a, i_b)$, where i_a is the identity presented in the survey question (i.e. teacher in "Who would you say is most likely to be *seen with* a teacher?") and i_b is a particular multiple choice answer of interest to us. For Study 1, we will estimate two different statistical models, one for each type of question, with the following form:

$$\begin{aligned}\phi(i_a, i_b) =& \beta_1 * semantic\_association(i_a, i_b) + \\ & \beta_2 * semantic\_similarity(i_a, i_b) + \\ & \beta_3 * affective\_similarity(i_a, i_b)\end{aligned} \tag{5.3}$$

Between the two types of survey questions, we will assume only that the values of the βs, the effects of these three different independent variables, differ. We therefore use Study 1 to explore our basic assumptions of how semantic similarity, semantic association and affective similarity impact the way survey respondents label individuals.

In Study 2, we will decompose effects of these three high-level variables into more detail. For example, in Study 1, we consider only the effect of semantic association on survey responses; in Study 2 we will estimate effects of identities' associations to particular institutional settings and the existence of counter-identity relationships. What we will find is that we can infer not only these βs but also infer details about *how* identities are semantically associated, i.e. what institutions they share, and *how* they are semantically related, i.e. what traits they share.

5.4 Data

A variety of publicly available datasets were used in this Chapter; in this section we provide an overview of each dataset used. We first review survey data used to measure semantic association, semantic similarity and affective meanings (which we use to compute affective similarity). We then discuss how we extract a consistent set of identities for further study that appear in all (or in some select cases, most) of these datasets.

5.4.1 Survey Data

Cue	Response Word	N. Respondents w/ Answer	Association Score
Man	Woman	99	0.66
Man	Boy	14	0.09
Man	Hole	3	0.02
Man	Huma	2	0.01
Man	Lady	2	0.01
Man	Moon	2	0.01
Man	Muscle	2	0.01
Man	Strong	2	0.01
Man	Wife	2	0.01

Table 5.2: Example set of responses given for the cue "Man" in the USF Free Association dataset

For a survey-based measure of semantic association, we use the USF free association dataset compiled by (Nelson et al., 2004). This dataset contains five thousand "cue" words that were given to at least ninety-four survey respondents (mean = 148). For each cue, respondents were asked to write the first word that came to mind that they felt was *associated* with the cue. As a

result, for each cue word we can construct a distribution of its association to other words based on the percentage of survey respondents that gave that word as an answer. Table 5.2 gives an example of the responses for one particular cue, "man", and the computed association strength determined simply by dividing the number of respondents giving a particular answer by the total number of respondents given this cue.

Word 1	**Word 2**	**SimLex-999 Similarity**	**Std. Dev. of Answers**
weird	odd	9.20	1.27
tiny	huge	0.60	0.59
communication	language	7.47	1.09
father	daughter	2.62	1.69
jar	bottle	7.83	1.44
book	information	5.00	1.13
wisdom	intelligence	7.47	1.33
dog	horse	2.38	1.39
modest	ashamed	2.65	1.00
hat	coat	2.67	1.76

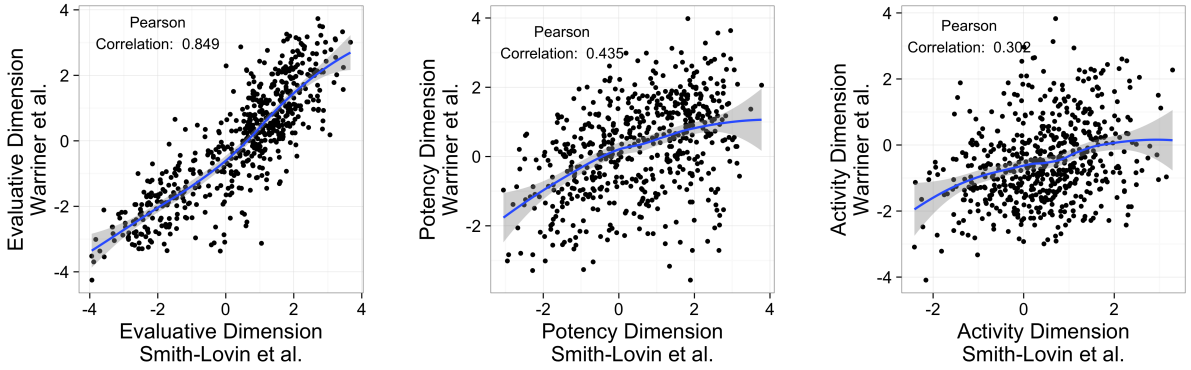
Table 5.3: Ten randomly sampled pairs of words from the Simlex-999 semantic similarity dataset, their similarity and the standard deviation of similarity scores given by respondents

For a survey-based measure of semantic similarity, we utilize the SimLex-999 dataset developed by Hill et al. (2014b) to test computational models developed to measure similarity. Hill et al. (2014b) pull 900 of the 72,000 possible pairs of cue-association words from the USF Free Association dataset (e.g. the pair man-woman from Table 5.2 to estimate similarity for. To this dataset, they add 99 pairs of words found in the USF Free Association dataset where each was either a cue word or a response word but that are not themselves associated. In total, then, 999 pairs of words are studied.

For each pair, Hill et al. (2014b) asked approximately 50 respondents to rate the similarity of the two words presented using a rating scale similar to those developed by Osgood (1969) and used by ACT researchers (Heise, 2007). Table 5.3 displays ten randomly sampled pairs of words given in the dataset and their similarity, which is measured on a scale of 1-10 and is assumed to be symmetric. For more details on the survey approach used by Hill et al. (2014b), we refer the reader to the original article.

For a survey-based measure of affective meanings of identities, we utilized three different dictionaries of EPA profiles. Two of the datasets were collected by ACT researchers- Smith-Lovin and Robinson (2015) collect mean affective ratings of 931 generic identities and Schröder et al. (2013) collect mean affective ratings for 56 social groups. In both cases, results are taken from individuals in multiple cultures, we only use data here from American survey respondents. The final dataset of affective meanings is drawn from Warriner et al. (2013), who collect EPA ratings of nearly 14,000 concepts using Mechanical Turk.

While the data collected by Schröder et al. (2013) is small enough that comparison to the other two datasets is difficult, we can compare words common between Smith-Lovin and Douglas (1992) and Warriner et al. (2013) to better understand how stable the mean affective ratings



(a) (b) (c)

Figure 5.1: Correlation between ratings of words shared by the Smith-Lovin and Robinson (2015) and Warriner et al. (2013) datasets. Figures a), b) and c) show correlations between the evaluative, potency and activity dimensions, respectively. A loess smooth line (blue) with 95% confidence intervals (grey) is also displayed

computed by these two datasets are. Further, both datasets have been used by a variety of previous works connecting EPA ratings to text (Ahothali and Hoey, 2015; Joseph et al.), making a comparison more interesting. Figures 5.1a), b) and c) show the correlation between the two datasets for all identities found in the Smith-Lovin and Robinson (2015) dataset that were also used in the Warriner et al. (2013) dataset (n=572). As is clear, correlation on the Evaluative dimension is quite high, whereas correlations along the other two dimensions are significant but also relatively low. This observation is consistent with prior work from respondents which suggests that variance of responses within a particular survey is generally much higher for the evaluative dimension than the potency or activity dimensions. In order to address differences in Figure 5.1, we take the mean of the ratings from both datasets as the final measure for any term appearing in both datasets.

5.4.2 Identity Selection

Both the survey and lexical resources we have described contain a significant number of words that are not identities, as well as far more identities than we would be able to use in a reasonably-sized survey. In this section, we describe how we narrow the scope of the present work to a small set of identities of interest, essentially ignoring all other information contained in these databases. From this larger subset, we draw subsets for further study in both Study 1 and Study 2.

To begin, we extract from the SimLex-999 dataset all those pairs of words for which we judged both words within the pair to be unambiguously representative of identities.[6]. In total, this represented 88 pairs of words for a total of 95 identities- obviously, certain words were used in multiple pairs. After discarding one of these pairs of words because there was no available data on the EPA profile of the word "victor", we are left with 87 pairs of words and 94 identities. We

[6]We did not consider the pair heroine-hero, as it appeared that the former term was interpreted as the drug rather than the female hero. We also ignored the terms god, devil and demon, judging them to be more representative of the religious concepts than their alternative identity meanings

then add to this set of 94 identities all of the identities in the Smith-Lovin and Robinson (2015) and Schröder et al. (2013) datasets. Words in this set are, for the most part, unambiguously used as identities and are therefore unlikely to be considered by our survey respondents as being representative of a non-person concept.[7]. We extract 966 unique identity labels from this dataset. Combined with the 95 identity labels from the SimLex-999 data, we are left with a total of 993 possible identities of interest (note that this is less than 966+95 because of overlap between the two datasets).

We then further subsetted this identity set to identities that were a) used most frequently as identity words, b) were in our lexical resources and c) were used relatively frequently. To ensure use as an identity, we first used both the NLP python library `spacy` () and Wordnet to identify any identity labels for which the dominant sense was a verb (e.g. suspect, accused) and removed these from consideration. To ensure that terms were in our lexical databases, we removed words which were not in Wordnet as a noun or an adjective and those that were not in the (fixed) vocabulary of the publicly available neural embedding models that we used. To ensure that identity words were used frequently, we check that they were used frequently in *either* a fairly informal medium, Twitter, or in a fairly formal medium, Wikipedia. To check the former case, we use the frequency counts of words from 56M tweets given by Owoputi et al. (2013) and retain only those identities used in more than 2500 tweets. To check Wikipedia, we first extract all 532,051 "clean" pages from a Wikipedia dump from December, 2015. A clean page is a page that is not labeled as a stub, that was still active one month after the dump was created, and that also has more than 50 views over 2 year span, where we pull one random hour for each day.

After this cleaning, we were left with 244 identities, a list that we then pared down based on pilot runs of our surveys where respondents had difficulties answering questions with particular identities.[8] To this list, we added a set of identities of interest to us in prior work on Ferguson[9] (Joseph et al., 2016b) and the Arab Spring[10] Joseph et al. (2016a) as well as Davenport's (2016) work on intersectionality[11] as well as two contemporary identities[12] and three complementary role identities[13] to those already in the dataset, leaving us with a total of 234 identities.

This set of 234 identities are used as a random "baseline" of identities in Study 1, as described below. Based on results from Study 1, we draw a subset of these identities to study further in Study 2.

[7]The following identities were removed from this set, as we judged them to have a dominant non-identity meaning:has been,in law,god,rat,dope,brain,doll, nut,authority,broad,tease,shrimp,star,fundraiser,grind,intimate, maverick, diner, nobody,divorce

[8]We removed the following identities: spy, infant, Chinese, deputy, german, visitor, Japanese, fellow, Russian, peer, VIP, bitch, female, gossip, supporter, runaway, superior, bastard, voter, maiden, angel, sucker, Hindu, tot, insider, client, editor citizen, Australian, pal, ally, European, fool, specialist, sweetheart, playboy, slave, beauty, dummy assistant, French, failure, devil, barber, technician, Irish, Korean, passenger, enemy, Polish, survivor, rival sidekick, British, server, person, gal, gamer, African, male, consumer, heel, Italian

[9]African-american,villain, police officer, defendant, gunman, juror

[10]extremist

[11]black man, black woman, white man, white woman, middle class, Hispanic, homosexual

[12]Trump supporter, Clinton supporter

[13]mom, sibling, Democrat

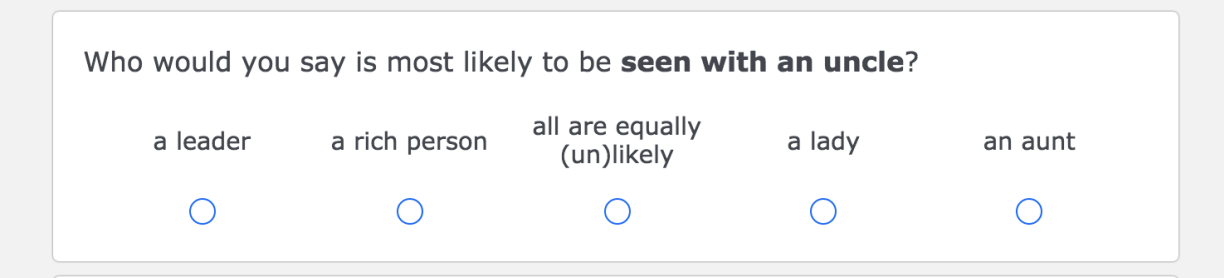


Figure 5.2: An example of a "SeenWith" question as seen by participants

5.5 Study 1: Exploring Effects of Similarity, Association and Affect on Identity Labeling

Our first study is geared towards understanding the impact of semantic association, semantic similarity and affective meaning on identity labeling. While intuition suggests that each should have an impact on different forms of the identity labeling problem, to the best of our knowledge no such proof exists. Here, we first discuss details of the survey, then provide some intuitions about expected results and finally an analysis of the survey data.

5.5.1 Survey Description

We begin with the set of 87 identity pairs for which we have pre-existing survey data on their semantic similarity (from the SimLex-999 data), semantic association (from the USF free association data) and affective profiles (from the EPA datasets). From here, assume that a particular pair contains identities A and B. We generate eighty randomized questions with this pair, twenty each from four types:

- **"IsA" A questions:** "Given that someone is a[n] **A**, what is **that same person** most likely to also be?"
- **"IsA" B questions:** "Given that someone is a[n] **B**, what is **that same person** most likely to also be?"
- **"SeenWith" A questions**: "Who would you say is most likely to be **seen with** a[n] **A**?
- **"SeenWith" B questions**: "Who would you say is most likely to be **seen with** a[n] **B**?

Each of these questions had five multiple choice answers. Within the answer set, the identity not in the question itself (i.e. B if A was in the question, or vice versa) was given as one of the answers. We then included three random identities from the set of 234 (that were not A or B) as alternative choices, along with the option "all answers are equally (un)likely" in order to allow respondents to opt out of answering questions they were uncomfortable with. An example question as seen on the survey is pictured in Figure 5.2.

Each respondent saw 40 random questions- with 80*87=6,960 questions to ask, we therefore required 174 survey respondents. Surveys were deployed on Amazon's Mechanical Turk to only "Masters"[14] and only those with IP addresses within the United States. To assess accuracy for respondents, we randomly sampled 5 questions from each respondent and ensured that answers

[14] b

did not appear to be entered randomly. No personally-identifiable information was collected, and all (anonymized) survey questions and responses will be made available.

5.5.2 Exploring the Hypothesis Space

	Semantic Similarity	Semantic Association	Affective Similarity	Examples	IsA	SeenWith
1				physician & doctor, friend & buddy		
2				student & pupil; teacher & instructor		
3				buddy & companion		
4				adversary & opponent; author & creator; champion & winner; leader & manager; politician & president		
5				wife & husband		
6				woman & man; child & adult		
7				doctor & professor; friend & mother; teacher & helper; brother & son		
8				adult & baby; author & reader; boy & partner; chief & mayor; dad & mother; daughter & kid; friend & guy; girl & maid; guy & partner; king & princess; lawyer & banker; man & child; man & sentry; man & warrior; pupil & president; sinner & saint; teacher & rabbi; winner & candidate		

Table 5.4: Examples of identities that are high (above 1 SD from the mean - colored red) or low (below the mean - colored blue) for each combination of high/low of semantic similarity, semantic association and affective similarity. We also provide predictions for the effect on the two types of questions (IsA questions and SeenWith questions). Red implies low odds of selecting the pair in the multiple choice questions, blue implies high odds, grey implies unclear odds.

Table 5.4 presents examples of identity pairs extracted from the SimLex-999 data that are high on some dimensions (blue cells) and low on others (red cells). Specifically, a pair is considered to have a high value of semantic similarity, semantic association or affective similarity if its value is one or more standard deviations above the mean, and to have a low value if it was below the mean. The difference in operationalization between "low" and "high" is intended to limit the number of examples produced; as Table 5.4 is for exploratory purposes, we can be relatively heuristic in these decisions.

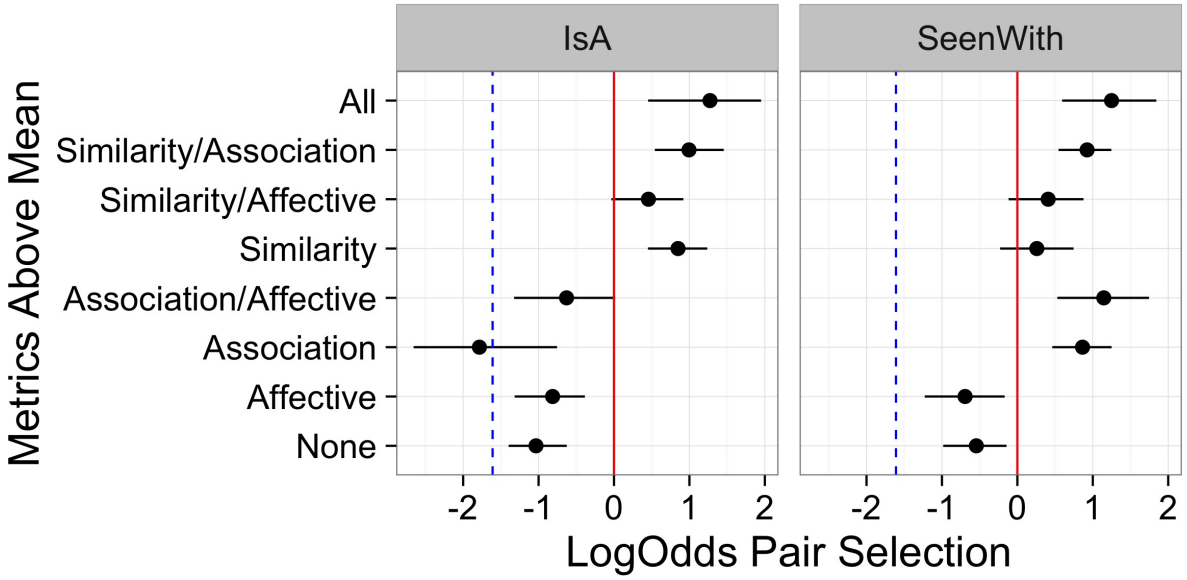


Figure 5.3: On the x-axis, the log odds of the identity not in the question being selected as opposed to the "All equal" option or any of the random identities presented. On the y-axis, identity pairs are broken into similar categories as in Table 5.4; see text for details. For each category, 95% bootstrapped Confidence Intervals are presented; vertical lines are drawn at a log-odds of 0 (red line; 50-50 chance of selection) and $log(\frac{1}{5})$ (blue dashed line; random chance of selection)

For each row of Table 5.4, we provide predictions based on the literature reviewed above of the odds of the identity in the multiple choice section of the survey questions being selected over the other random answers presented. We will refer to these odds as the *odds of selection* throughout this section. In Figure 5.4, blue represents high odds, red low odds, and grey represents an unclear picture given the prior literature. When all measures are high (row 1), we should expect identity pairs to have high odds of selection regardless of the type of question posed - physicians and doctors are synonyms, so we should expect the same individual could have both identities and be seen with others of the opposing identity. The converse is also true - when there is no relationship between two identities (row 8), we should expect that odds of selection are relatively low for both types of questions.

In all other cases, however, competing factors suggest an unclear picture of results. This is true even if we assume that semantic similarity impacts only "IsA", questions and that semantic associations impact only "SeenWith" questions, because we expect affective similarity may impact both types of questions. In this case, only when a semantic dimension agrees with affective similarity can we be certain about our predictions. We turn now to an analysis of survey results to better understand these cases and to confirm or deny the predictions we were able to make in Table 5.4.

5.5.3 Results

The y-axis of Figure 5.3 presents eight classes of identity pairs differentiated by which of the three metrics they were above-average for. These are roughly the same eight rows (and are in

the same order) as those presented in Table 5.4. In Table 5.4, however, we show examples of identities far above (1 SD) the mean on particular categories and below the mean on others. Figure 5.3 instead groups identities only on whether or not they were above or below the mean on each category in order to include all pairs of identities surveyed and thus be able to compute better confidence intervals. Although Figure 5.3 and Table 5.4 are therefore clustered slightly differently, we can still use Figure 5.3 as a rough guideline to how well survey results matched our expectations.

The x-axis of Figure 5.3 shows the log-odds of selection. This value measures the proportion of times the identity pair element in the answer set was selected out of the 20 randomized questions generated for that question type and that arrangement of identities. So, for example, if "woman" were selected 19 out of 20 times when given as a possible answer for the question "Who would you say is most likely to be seen with a man?", then the log-odds of selection for the "man-woman" pair would be $\frac{19+1}{1+1}$, where a $+1$ is added to avoid zero-valued denominators. Justification for the use of this value can be found in Appendix B. Finally, Figure 5.3 also displays two baselines to consider- a red, solid line is drawn at a log-odds of 0, representing the odds of the identity being selected as the answer more often than not. The blue, dashed line is drawn at a log-odds of 20%, that is, the odds of the identity being selected more often than random.

Figure 5.3 shows that when identities are above average on all metrics (top row in Figure 5.3, row 1 in Table 5.4), log-odds of selection were indeed high. More generally, Figure 5.3 appears to confirm that high semantic similarity breeds high log-odds of selection for "IsA" questions regardless of the other metrics, and high association breeds high log-odds of selection for "Seen-With" questions.

The remaining factor, affective similarity, seems to have little independent effect on how respondents answered questions. Specifically, taking the confidence intervals of the bottom two rows of Figure 5.3 as indicators, little evidence exists that affect alone changes the way we label individuals in any significant way over the case where no strong semantic or affective relationship exists. The only place in which a possibly interesting effect of affect can be seen to exist is in "IsA" questions, when affect in addition to association seems to increase the log-odds by a considerable amount. We return to this point below.

When all three factors are below average (last row in the figure/row 8 in the Table), the log-odds of selection are low. The log-odds are still, however, noticeably greater than pure chance. This is likely due to the way that word pairings were selected in the SimLex-999 dataset-Hill et al. (2014a) sampled from existing cue/response pairs in the USF free association data. Consequently, we work here with identity pairings that largely have some form of association, and thus their relationship is stronger than a random baseline in almost all cases. Even with this selection bias, however, the log-odds of selection when "IsA" questions are asked of identity pairs that are only above average on association (the "Association" row in Figure 5.4, row 6 in Table 5.4) are not statistically different from random selection. These pairs, which we identified as likely to be be counter-identity pairs above, thus are distinguished in an important way from other types of identity pairs.

Figure 5.4 furthers these observations. In the figure, we show two subplots, one each for the two different types of questions. Within each subplot, each of the 87 identity pairs studied is given by two points, one each depending on which identity was shown in the question and which was given as a possible answer. The x-axis again displays the (scaled) semantic similarity

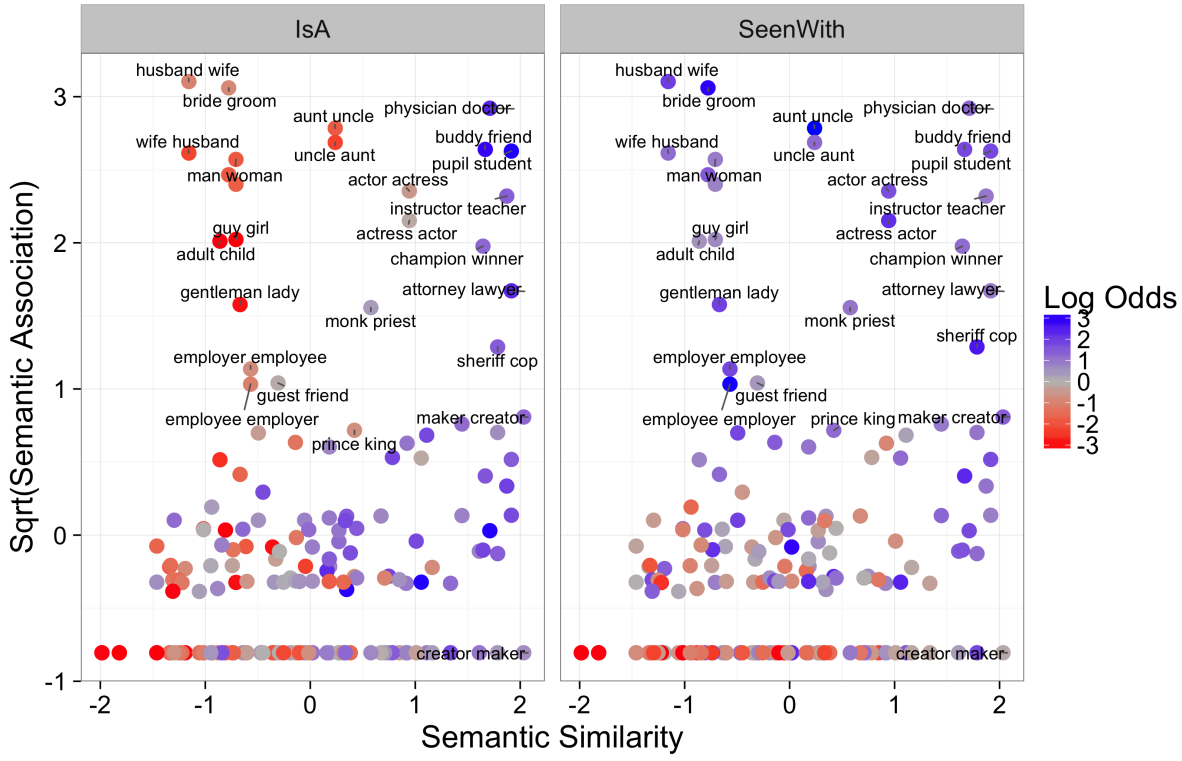


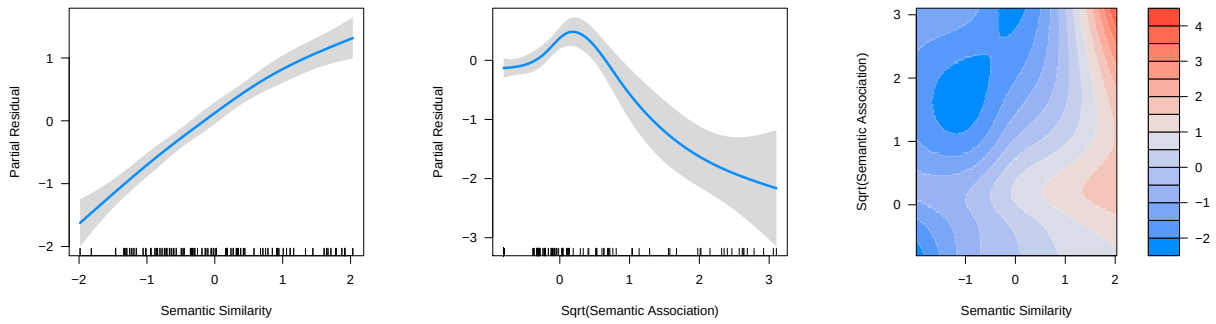
Figure 5.4: Results for the two different types of questions for log-odds (represented by color), semantic association and semantic similarity. Within each subplot, each identity pair is shown by two points, one each depending on which identity was shown in the question and which was given as a possible answer. Outlier points are labeled based on low-probability with an overlying density estimator

of the identity pair, the y-axis now displays the (scaled) square root of semantic association.[15] Note that these zero-associations are a result of our use of both "directions" of each identity pair. Thus, while we are guaranteed some non-zero association in most of the pairs collected by Hill et al. (2014a) in one "direction", in the other there is no such guarantee. Finally, each point is colored in Figure 5.4 by the log-odds of selection - the darker the blue, the higher the log-odds, the darker the red, the lower the log-odds.

Figure 5.4 suggests that more semantic similarity entails higher log-odds of selection in "IsA" questions and more semantic association leads to a higher log-odds of selection in "SeenWith" questions. It also suggests that identities in the top-left of the plot, high in association but low in similarity, have low log-odds of selection in the IsA question set and very high odds of selection in the SeenWith question set. The labels on these pairs allow us to recognize them as counter-identity pairs.

We can further confirm these results via a more rigorous statistical model. To do so, we fit a binomial generalized additive model (GAM) (Hastie and Tibshirani, 1990) using the `mgcv` pack-

[15] We use the square root as it better represents a strong distinction between a zero-valued association and a small but non-zero association.



(a) (b) (c)

Figure 5.5: Results from a GAM fit to logit of the odds of selection for "IsA" questions. Figures a) and b) show fit lines (blue bar) and 95% confidence intervals of the fit (grey shadows) for semantic similarity and semantic association, respectively. Figure c) shows fit with the interaction term, where color represents log-odds of selection.

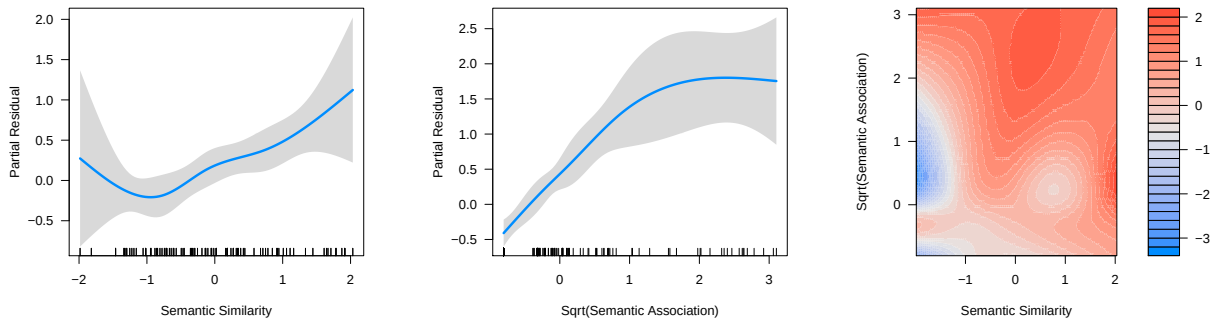
age in R (R Core Team, 2015; Wood, 2006) to results on each type of question independently.[16] In essence, generalized additive models are generalized linear models that relax the assumption of linearity, instead fitting a (possibly multi-dimensional) curve to each model parameter. The "wigglyness" of these curves, or functions, is controlled by some form of penalty; in the `mgcv` package, this penalty is determined via cross-validation. This penalty helps to ensure that the fitted functional form does not over-fit the training data.

The model we use predicts the logit of the odds of selection (in other words, a logistic regression on the odds of selection) by fitting tensor product bases on semantic similarity and semantic association independently as well as a multivariate tensor basis on their interaction. Figure 5.5a) shows the fit to semantic similarity to partial residuals of the logit odds of selection for IsA questions only. Figure 5.5b) shows the same for semantic association, and Figure 5.5c) shows a heatmap of the interaction effect. In Figure 5.5a) and Figure 5.5b) partial residuals on the y-axis essentially show the fit of each variable after removing effects of the other variable and their interaction.

Figure 5.5a) shows that, controlling for the interaction effects of the two variables, semantic similarity has an almost perfectly linear relationship with log-odds of selection in IsA questions. Semantic association, net of semantic similarity, meanwhile, has a curvilinear effect showing that when association is high and similarity is low, there is a strong negative effect of association, but that a zero-association leads to less overall similarity. Thus, there exists a sort of acceptable region of association for "IsA" quesitons. Note that within this region, the relationship is relatively linear as well. Figure 5.6 provides the same model except fit on data from SeenWith question responses. Here, we observe that semantic association has, as expected, a strong impact on log-odds of selection. We also observe, however, that net of semantic association and the interaction term, semantic similarity has a significant effect on log-odds of selection.

Figure 5.5 and Figure 5.6 thus provide additional confirmation of our intuitions, as well as

[16] Chapter 8 and Chapter 9 of Shalizi (2013) provide a nice introduction to GAMs and tensor product bases.



(a) (b) (c)

Figure 5.6: The same GAM model as in Figure 5.5, except here we fit to data from only "SeenWith" questions

providing the novel insight that even when identities are semantically disassociated, underlying similarity (e.g. in the case of adversary and opponent) can impact our perception of which identities are likely to be seen together. However, the statistical models we fit serve two more important purposes as well. First, while the models provide reasonably strong fit to the data (explaining 49.7% and 38.1% of the residual variance in the "IsA" and "SeenWith" models, respectively), a significant amount of variance remains.

Of course, one possible explanation for this is that affective similarity is not accounted for in the model. One way to consider evidence that affective similarity indeed explains some of this variance is by attempting to fit it to the residuals generated from these models. Flaxman et al. (2015) describes a process to do model such partial correlations by correlating residuals after a whitening step with a Gaussian Process model. We use a similar approach, attempting to fit a generalized additive model of affective similarity on the residuals of the above models. When doing so, we find there is no significant effect of affective similarity on residuals, providing strong statistical evidence that affective similarity had little or no effect on how respondents answered questions net of semantic similarities and associations.

Second, we were able to observe the shape of these effects - results suggest that in controlling for interaction between the two variables, large regions of the resulting effects are nearly linear for semantic similarity and on the square root of semantic association. Even in Figure 5.5b), where the effect is most obviously non-linear, variance in residuals from the non-linear fit is high enough that a linear approximation would likely capture much of variance captured by the additive model. This observation will be useful in the following study.

Finally, these figures show large standard errors for the fit lines at the ends of the distribution, suggesting the model struggled with outliers. Table 5.5 shows the top ten cases in which the "SeenWith" model under-predicted the true log-odds of selection. The table presents some surprising scores for semantic association - for example, "king" and "princess", as well as "king" and "prince", are both less associated than the average identity pair in our dataset. Given that these are drawn from a similar domain, these numbers are surprising at first sight.

The cause of this is, we believe, the use of free association as a metric for association. The

Rank	Query Term	Possible Answer	Predicted Log-odds	Log-odds of Selection	Scaled Semantic Association (sqrt)
1	captain	sailor	-0.40	2.35	-0.80
2	sailor	captain	0.33	3.00	-0.08
3	author	reader	-1.28	0.98	-0.80
4	worker	employer	0.23	2.40	-0.32
5	king	princess	-0.38	1.79	-0.80
6	princess	king	0.09	2.23	-0.10
7	king	prince	0.30	2.40	-0.28
8	employee	employer	1.14	3.22	1.03
9	professor	student	-0.13	1.85	-0.31
10	president	politician	0.52	2.48	-0.32

Table 5.5: Top ten identity pairs for the "SeenWith" model in terms of under-predicted by the model relative to the true log-odds of selection

problem with using free association is that a single association can "eat up" a significant amount of the semantic association in a free association task, masking minor but still important associations. Specific to the case of "king", the counter-role queen takes most of the association score in a free association task, meaning other identities within the same institutional structure that are still highly associated are given lower scores than would be expected. A related example is the identity mother, which has a high free association score to "father" but no association with "dad".

Predictions for our "SeenWith" model are thus hindered by the specific way in which we measure semantic association. The same can be said for the results of the "IsA" model - more specifically, the measurement assumption that semantic similarity is symmetric leads to difficulties in prediction. Table 5.6 presents ten identity pairs where log-odds of selection differed the most depending on which identity was presented in the question. As pairs had the same value for semantic similarity regardless of which identity was presented first, these pairs represent obvious cases where the model would be unable to capture variance in the data. They also present obvious cases where semantic similarity cannot be assumed to be symmetric -for example, a "president" tends to be a "politician", but a politician is not always a president. These asymmetries are due to the naturally occurring hierarchy of identities, and emphasize the variety of ways in which identities can be considered to be similar.

5.5.4 Discussion of Findings

Results in Section 5.5.3 used a variety of approaches to suggest two unexpected effects:

- Affective similarity had no demonstrable effect on how respondents answered questions when controlling for semantic factors
- Semantic similarity has a significant effect on log-odds of selection in "SeenWith" questions, even with stringent controls mediating the effect of association and any interaction effect with association

With respect to the former point, at least four possible explanations exist. First, in fitting

Rank	Identity 1	Identity 2	Log-odds, Iden. 1 in question	Log-odds, Iden. 2 in question
1	stud	guy	-1.61	1.73
2	president	politician	-0.17	3.14
3	princess	bride	-2.48	0.41
4	worker	employer	0.98	-1.85
5	warrior	man	-2.08	0.56
6	teacher	rabbi	0.15	-2.25
7	mayor	chief	-0.08	-2.40
8	manager	leader	-0.37	1.85
9	baby	adult	-3.09	-0.89
10	worker	mechanic	1.22	-0.76

Table 5.6: Top ten identity pairs in terms of difference in log-odds of selection in "IsA" questions depending on which identity was presented in the question (vs. as a possible answer)

affective similarity to residuals of these semantic effects, we made an implicit causal assumption which may have led us to underestimate affective similarity's impact. However, the weak correlation between affective similarity and log-odds of selection in general suggest this to be an appropriate modeling decision. Second, affective profiles of identities may simply be just another element of the static traits assumed to drive decisions on semantic similarity; consequently similarity should be expected to eat up any variance explained by affective similarity. Third, it is possible that the identity pairs in our dataset were biased towards those for which affect was not a factor; however, one could imagine identity pairs (e.g. "moron" and "idiot") that were linked primarily via affect. Finally, it may be the case that affective similarity simply does not have a strong effect on the identity labeling process as tested here.

The relationship of semantic similarity to "SeenWith" questions suggests that a strong distinction between dynamic properties, static traits and shared institutional settings is difficult, especially in the context of identity labeling. The complexity of this distinction is further increased by two inadequacies in the data we use for semantic similarity. First, Hill et al. (2014a) assume a symmetric relationship exists for semantic similarity. Our survey results indicate that for identities, this assumption struggles with certain asymmetric, hierarchical relationships in defining similarity.

Second, the assumption suggests that similarity can be measured unidimensionally. Our results suggest that at least two forms of similarity played a role in how respondents answered questions. Similarity as typically defined, and thus measured, by computational linguists tends to represent these taxonomic relationships, as in, "a lawyer `isA` professional". The effects of such hierarchical, denotive relationships has already been discussed. However, with respect to identity, similarity also refers to labels that may apply to the same individual regardless of taxonomic relationship - in sociological terms, the extent to which two identities are cross-cutting (Blau, 1977). Cross-cutting similarities are unlikely to be captured via denotatively organized data sets like WordNet, or even, it seems, from explicit questions about semantic similarity.

Where they do seem to be captured, however, is in the survey methods presented in this chap-

ter. A good example is the identity pair "secretary-woman", which had a log-odds of selection of 1.22 (sixteen out of twenty) and a scaled semantic similarity of -1.29 (1.29 standard deviations below the mean). These two identities have little, if any, denotive relationships, and it seems that when the question of similarity is posed explicitly as by Hill et al. (2014a), respondents were focused on this connection. In contrast, via the more subtle mechanisms utilized used in our survey, we are able to uncover the well-known gender bias towards considering secretaries as being primarily women.

This finding is comparable to the observations of Vaisey (2014). While we do not necessarily condone the dual-process modeling perspective, it seems reasonable to assume that cross-cutting similarity could roughly be underscored by subconscious stereotypes that elude denotive, conscious categorizations that are elicited via explicit questions about similarity. Regardless, inferring these types of relationships in a principled fashion is a chief goal of Study 2.

Distinctions of counter-identity pairs also sometimes eluded denotive characterizations - many of the identity pairs that fit this typology of survey responses were not categorized in Wordnet as antonyms. Identity/counter-identity pairs also impacted our measurement model of semantic association, overshadowing smaller but still relevant semantic associations between identities.

Despite these measurement issues for semantic similarity and semantic association, three main intuitions presented in Section 5.5.2 did seem to hold:

- Semantic similarity appears to be a powerful driver in how survey respondents answered our "IsA", or multiple identity questions
- Semantic association appears to be a powerful driver in how survey respondents answered our "SeenWith", or same-context questions
- Counter-identity relationships show distinct answer patterns - they had high log-odds of selection in "SeenWith" questions but low log-odds in "IsA" questions.

Figure 5.7 further underscores this last point. We plot the log-odds of selection in the two different types of questions for each identity pair. We then cluster this data using a Gaussian mixture model[17] and determine the range of strongly-fitting models over a range of possible numbers of clusters and model assumptions via BIC. We find that a four-cluster model fits reasonably well to the data, with a three-cluster model fitting slightly better but visually performing poorly on outliers.

In the figure, cyan-colored identities (bottom left) were essentially "incomparable" in that survey responses were essentially random for these entries. Respondents found them to be neither similar nor associated. The set of identities in the top right (red) were highly associated and highly similar - sociologically speaking, they represented similar identities within similar social contexts. Green points (top left), on the other hand, represented identities found in the same context but that had significantly different properties - counteridentity pairs. Finally, purple identities captured those that were difficult to classify and also a small set of identities that seemed to share properties but were not often assumed to share social contexts. An example is the point in the lower right of the plot, representing the identity pair "man - husband".

Figure 5.7 thus suggests that one can characterize the identity pairs studied in this section into four types, re-structured from the eight types in Table 5.4:

[17]Using the `mclust` package in R (Fraley et al., 2012)

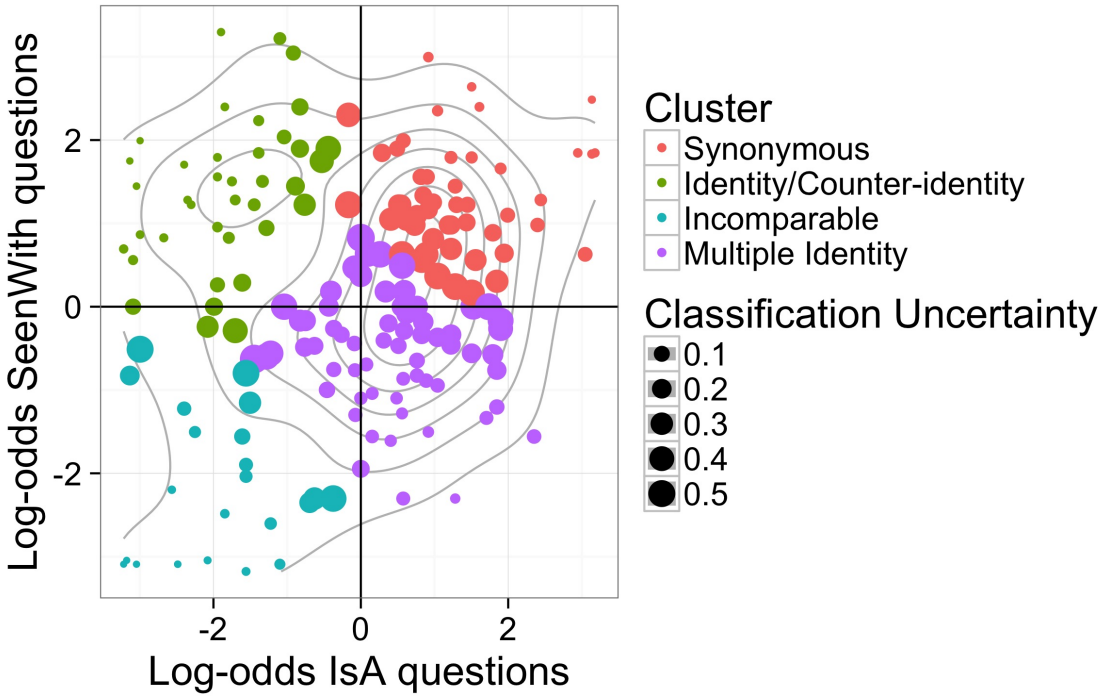


Figure 5.7: Log-odds of selection for IsA questions (x-axis) vs. SeenWith questions (y-axis) for all identity pairs. Points are colored by association with a particular cluster in a Gaussian mixture model and sized by uncertainty (1-probability of most likely cluster). Clusters are named, explanations are given in the text. A line with slope=1 and a density estimator are added for reference.

- **Counter-identity pairs** appear frequently together in the same social contexts but have very different properties on which similarity is judged; i.e. husband and wife.
- **Synonymous identity pairs** refer to the same concept, and thus are highly likely to co-exist within a particular individual and across individuals within a particular setting; i.e. physician and doctor
- **Incomparable identity pairs** represent those pairs that are essentially random - there is little chance we would ever expect the same individual to share both identities or to see these identities together in any context; i.e. dentist and colnel
- **Multiple identity pairs** counter-intuitively represent identities rarely expected in the same context but that tend to represent the same individual

This typology uncovered from the survey data suggests that identity pairs have important and distinct kinds of relationships that lead respondents to answer IsA and SeenWith questions differently. In this section, we showed convincing evidence that these differences can be accounted for to some extent by the amount of semantic similarity and the semantic association between these two identities. In the following section, we consider further evidence for the findings presented here via the use of a second survey.

Type	Sub-type	Identities
Institutional Setting	Politics	conservative, Democrat, liberal, Republican, politician, senator
Institutional Setting	Family	brother, sister, daughter, son, father, mother
Institutional Setting	Law	judge, criminal, lawyer, witness, cop, police officer
Institutional Setting	Medicine	doctor, physician, surgeon, nurse, patient, dentist
Institutional Setting	Business	executive, consultant, secretary, intern, banker, boss
Social Categorization	Gender	woman, guy, girl, boy, man, lady
Social Categorization	Age	teenager, kid, child, toddler, adult, minor
Social Categorization	Race/Ethnicity	black, white, Hispanic, Asian, Arab, American
Affect	Negative Affect	thug, idiot, jerk, goon, punk, bully
Random		principal, scientist, coach

Table 5.7: Identities used in Study 2 and the "institutional setting" they are drawn from. Details are provided in the text

5.6 Study 2: Exploring How Identities are Similar/Associated

Survey 1 presented interesting findings about how respondents used (or did not use) semantic similarity, semantic association and affective similarity to answer our identity labeling questions. We also observed that identity relationships could be characterized into one of four general types - incomparable identities, culturally synonymous identities, multiple identity pairs and counter identity pairs.

However, the survey also had several shortcomings. Perhaps most importantly, we used a set of identity pairings drawn by Hill et al. (2014a) that may under-represent the different types of relationships amongst identities and may also have been biased against observing affective similarity as an effect.

In Study 2, we attempt to address this selection bias by using an expanded set of pairs sampled in a more uniform fashion. From the set of 234 identities described in Section 5.4.2, we draw the 57 identities shown in Table 5.7 from three different generic types of mechanisms traditionally used to group identities together. We draw six identities each from five institutional settings described by Heise and MacKinnon (2010) and implemented in Interact Heise (2007). We then include three sets, also of six identities each, from three traditional forms of social categorization - gender, age and race/ethnicity. We also include a set of six "negative affect" terms, related only by the fact that they share similar affective meanings. Finally, we include three random identities as a mechanism for comparison.

5.6.1 Survey Details

From the 57 identities in Table 5.7, we create survey questions as follows: for a given identity, we generate 14 random sets of the 56 other identities; each set contains four identities. We then generate one "IsA" and one "SeenWith" question for each of these sets, where these four identities constitute the possible answers to the question, and the given identity is used in the question text. This process is then repeated ten times for each identity. This process generates

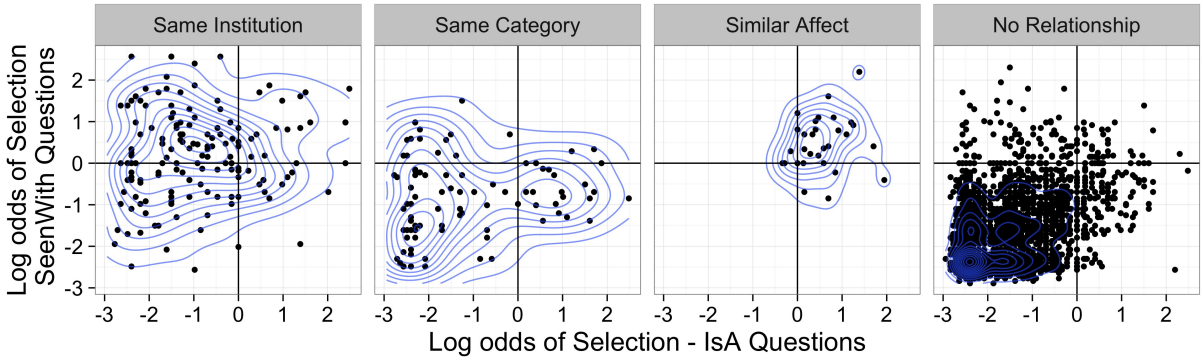


Figure 5.8: Log-odds of selection for SeenWith questions (y-axis) and IsA questions (x-axis) for each identity pair considered in Survey 2. Four different subplots are shown; one each for the three different types of relationships induced by our selection of identity labels and one for all identity pairs which were not in the same selection group in Table 5.7. A density estimator is overlaid in blue for each subplot

exactly ten questions for each of the 3,192[18] identity pairs we study for each type of question. The intention was therefore to have exactly ten questions for each identity pair for each question type where the first identity in the pair is shown in the question and the second identity in the pair is shown as a possible answer. In each case, the other possible answers were randomly selected.

Unfortunately, at least one of the surveys we developed suffered from a since confirmed bug on Qualtrics where the option to present questions an even number of times fails in unclear cases. Due to this error, some of our questions - 3.4% were asked less than 6 times, 40% were asked less than ten times, and 40.4% were asked more than ten times. While this did not greatly effect our analyses, which relied on latent space representations of our identities, it is important to note for purposes of any future work with the dataset.

Such issues aside, the process described generated 15,960 questions, which were split evenly, 40 questions per respondent, between 399 Mechanical Turkers who were located in the United States, had greater than a 95% completion rate and who had completed over 1,000 HITs. Responses were checked for bad survey responses in the same fashion as in Study 1. We now move to a discussion of the statistical model we will use to analyze the data, our expected results and finally on to how these expected results matched with the data.

5.6.2 Survey Results

Figure 5.8 displays log-odds of selection for IsA questions and SeenWith questions for each of the 3,192 identity pairs considered in Study 2. The left-most subplot shows pairs that were not grouped a priori by the selection criterion for the survey -that is, they were not in the same row in Table 5.7. The other three subplots show identity pairs that were grouped together a priori for this study in Table 5.7.

Figure 5.8 shows results in Study 2 depart in two important ways from results in Study 1. First, the identities that we utilized which we determined to be affectively similar were largely representative of the identities that were culturally synonymous. In Section 5.5.4 we discussed four possible reasons for this: modeling decisions, selection bias, incorporation of affective sim-

[18] 57*56=3,192

Rank	Top 10 Log-Odds IsA	Top 10 Log-Odds SeenWith
1	lawyer - adult	mother - toddler
2	criminal - thug	principal - child
3	boy - son	mother - child
4	physician - adult	girl - mother
5	kid - brother	toddler - mother
6	cop - guy	mother - kid
7	woman - daughter	child - mother
8	secretary - lady	lady - child
9	woman - mother	thug - criminal
10	nurse - adult	boy - mother

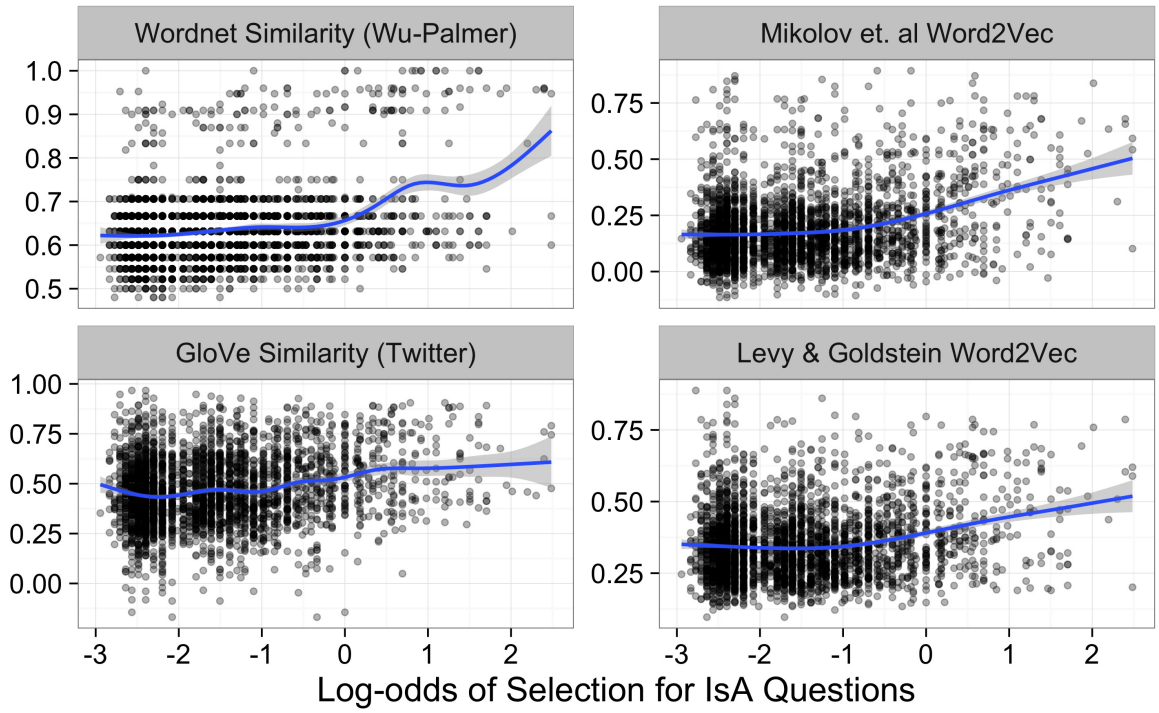
Table 5.8: Top 10 identities in log-odds of selection for IsA questions (left column) and SeenWith questions (right column) that had no a priori assumed relationship.

ilarity into measurements of semantic similarity, and a true lack of an effect. Figure 5.8 gives strong evidence that the latter option is incorrect. It therefore seems likely that some combination of selection bias, modeling decisions and the relationship between semantic and affective similarity may have resulted in the null effect of affective similarity on identity labeling in Study 1. How best to tease out these differences will be an important focus of future work.

The second difference we observe between Study 1 and Study 2 is that by and large, most identity pairs are incomparable. This is a result of using a sample of identities for Study 2 that were less biased towards an existing association than the identity pairs found in the SimLex-999 dataset, we see that Thus, regardless of the labeling task a person is presented with, information about another identity existing within the context or that has been applied to an individual greatly reduces the possible universe of other labels we are likely to apply.

Beyond these two differences between survey results observable from Figure 5.8, we also see that results in Study 2 are strongly suggestive of the four types of identity relationships discussed at the end of Section 5.5.4 are also observed in Study 2. Moving from the left-most subplot to the left-most subplot in Figure 5.8, we see first that identities pairs within the same institutional setting were frequently highly associated but minimally similar, confirming the existence of counter-identity relationships for many identities that share an institutional settings. The second sub-plot shows that identities sharing only the same category were largely disassociated, and were then either highly similar or dissimilar depending on if they shared that property (e.g. “kid” and “toddler”) or represented opposite ends of that category (“kid” and “adult”). The third subplot shows that identities that had similar affective profiles were largely representative of they synonymy relationship - they were both highly associated and highly similar. Finally, we see that identities that crossed the boundaries outlined in Table 5.7 were largely incomparable - they were rarely found to be either associated or similar.

The results in Figure 5.8 thus show a nice replication of the four-category depiction of identity relationships we discussed in Section 5.5.4. However, it is also clear that the separation induced by Table 5.7 does not capture all of the counteridentity, synonymous or multiple identity pairs in Study 2. In other words, this heuristic split of the data is unable to capture even simple



. Loess smooth lines are plotted in plue

Figure 5.9: Four commonly used metrics of semantic similarity derived from linguistic resources (each subplot) and how they correlate with log-odds of selection for IsA questions for identity pairs in Study 2 (x-axis of each plot)

interrelationships, e.g. those between identities within particular institutions that are representative of social categories. This is perhaps best shown by example - Table 5.8 shows the top ten identities in log-odds of selection for IsA and SeenWith questions that had no relationship based on the split in Table 5.8.

These identity pairs represent intuitive examples of the ways in which the binary categorization scheme of Table 5.7 fails to capture underlying associations and similarities between identities. Unfortunately, we have no survey-based measurements to confirm or deny this underlying similarity or association. Further, measurements, particularly of semantic similarity, from computational linguists do not seem to model this process well. Figure 5.9 presents the correlation between the log-odds of selection for IsA questions, which we would expect to correlate strongly with semantic similarity, with output from four models purported to measure semantic similarity using either a knowledge-based or corpus-based approach. Clockwise from the top left, these models are a corpus-based approach using Wordnet (Wu and Palmer, 1994), the original Word2Vec embeddings from Mikolov et al. (2013a), the GloVe embeddings trained on Twitter data from Pennington et al. (2014) and the dependency-parse incorporating Word2Vec model from Levy and Goldberg (2014).

These models seem to capture obvious semantic similarities (i.e. there is a correlation when these metrics are high), but in general fail to capture semantic similarity as it is relevant to our

Parameter	Interpretation
I	universe of all possible identity labels
S	The set of all institutional settings being modeled
P	The set of all dynamic characteristics, static traits and external cues being modeled
i_y	The identity applied to a person y
c	*Observable* features of a context in which an identity labeling occurs - for example, the institutional setting
$\phi(i, y, c)$	A function used to determine how likely a particular identity i is for particular individual y in a particular context c
θ_i	A vector representing the "attachment" of identity i to each institutional setting modeled
$\mu_{i,e}$	The $e$valuative dimension of the sentiment profile for identity i_m. μ_{i_m} is a three-dimensional vector representing the EPA profile of i_m
τ_i	A continuously valued vector of length $\|P\|$ representing the strength of the relationship between an identity i and each element of P

Table 5.9: An overview of the variables used in this section

problem. As we have discussed in the context of Hill et al. (2014a), this stems in part from the fact that these models are optimizing a different notion of similarity than is relevant for the study of identity. Even, however, if these models generated similarity scores that were aligned with similarity as we have defined it here, they would be of little use in developing a parsimonious representation of *how* identities are similar. We now turn to the development of an identity labeling model that helps to accomplish this task.

5.7 A Concrete Mathematical Model of Identity Labeling and its Application to Survey Data

In this section, I begin by describing a stronger model for identity labeling that can be used to better understand how identities are similar and associated, among other uses. I then introduce how this model can be used to form a likelihood function for the survey data in Study 2. To do so, I will model the ways in which we expect the identities in Table 5.7 to be similar and/or associated. I can then infer the parameters of this model from the survey data itself. Instead of using semantic similarity and semantic association to understand how people label others, we thus use how people label others to infer semantic similarities and associations between identities. I finish this section with an exploration of initial results utilizing this model to learn more about how identities are similar and associated.

Table 5.9 displays the variables used in this section with a brief description - note that many of these also appear in Table 5.4, we copy them here for convenience. Also, note that the assumptions developed in Section 5.3 apply here as well.

5.7.1 A Concrete Mathematical Model of Identity Labeling

I begin by further formalizing generic representations of semantic association, semantic similarity and affective similarity. I then use these equations to define a new ϕ. I then talk about how one can operationalize this ϕ in a way that is identifiable and useful for a particular set of identities utilizing the concept of the counter-identity from Burke (1980).

Formalizing Semantic Associations

From Heise and MacKinnon (2010), we known that one strong determinant of semantic association between two identities is the extent to which they are utilized in common institutional settings. We can express the expected amount of semantic association between two identities i_y, the identity of y whom we are trying to label, and another identity available in the context, say i_c, as:

$$semantic\_association(i_y, i_c) = (\theta_{i_y} - \theta_{i_c})^2 \tag{5.4}$$

In Equation 5.4, θ_{i_y} is a vector of length $|S|$ representing the "activation" or connection of i_y to each institutional setting in S. For example, if i_y were to be "student", it would likely have high values for the dimension (column) of θ representing the "schoolyard" setting, and low values for the dimension of θ representing the military setting. The matrix Σ_θ is an $|S|$x$|S|$ diagonal matrix of weights representing the importance of each setting in semantic association - perhaps, for example, the fact that two identities appear in the family setting makes them more highly associated than the fact that they both appear in the hospital setting.

Formalizing Semantic Similarity

Let us again assume that x already knows that y has a particular identity $i_{y'}$, and is interested in applying a second label to that individual. This situation is analogous to a variety of other research problems in this area - for example, work on implicit bias has sought to understand how likely it is that a woman is also labeled as a potential employee (Greenwald and Banaji, 1995). In our case, for example, x may be informed that $i_{y'}$ is a child, and is then asked what else y would be likely to be by giving another label i_y.

From Section 5.2, we know that the extent to which two identities are semantically similar is largely based on some function of the dynamic characteristics they share (Burke, 1980) and the similar connections to static traits and external cues they have (Freeman and Ambady, 2011) that cause them to be highly activated by similar external signals. These properties may, of course, be latent, but the literature provides a fairly strong set of candidate properties that are of high importance, specifically gender, age and race.We can represent this by giving each identity a vector of properties, τ, and considering similarity as an aggregate metric over these properties. Formally, then, we can state:

$$semantic\_similarity(i_y, i'_y) = (\tau_{i_y} - \tau_{i_{y'}})^2 \tag{5.5}$$

Here, τ is a vector of length $|P|$, where P represents the set of all dynamic characteristics, static traits and external cues that influence decisions of how similar two identities are. Again, Σ_τ is a matrix of weights on these different properties. For example, gender might be more important than age in determining similarity.

Formalizing Affective Meaning

While the prior section suggested that affective similarity was important in identity labeling, we retain it here in order to further test this idea. We show in this section that when tested directly, affect does indeed seem to have an effect. The affective meaning of identities, their behaviors and the affective properties of the setting in which they are placed have been shown to be predictive of both the multiple labels we apply to a particular individual and how we assign labels to individuals within a particular context (Heise, 2007). Therefore, it is useful to generally express the affective similarity of two identities as simply a function of two general identities, i_a and i_b:

$$affective\_similarity(i_a, i_b) = (\mu_{i_a} - \mu_{i_b})^2 \tag{5.6}$$

Here, μ_{i_a} is a vector of length 3 composed of the mean affective profile of i_a - i.e. its value on the evaluative, potency and activity dimensions of affect purported by Affect Control theorists.

Defining ϕ

We again will define ϕ as a pairwise function $\phi(i_y, y, c)$. As before, we will also model differences between the two types of questions in the survey by simply changing the weights of the importance based on the type of question being asked. For simplicity, we here do so by defining two different functions, one for each type of question. Note, however, that they will differ only in the weighting applied and when they are used, not in the underlying meanings they attribute to identities. First, $\phi_c(i_y, i_c)$ is used when an identity, i_c is given in the same context as y (i.e. "Who would you say is most likely to be *seen with* a teacher?"):

$$\begin{aligned}\phi_c(i_y, i_c) &= \beta_1 * semantic\_association(i_y, i_c) + \\ &\quad \beta_2 * semantic\_similarity(i_y, i_c) + \\ &\quad \beta_3 * affective\_similarity(i_y, i_c) \\ &= -\Big(\beta'_{\theta,c}(\theta_{i_y} - \theta_{i_c})^2 + \beta'_{\tau,c}(\tau_{i_y} - \tau_{i_c})^2 + \beta'_{mu,c}(\mu_{i_y} - \mu_{i_c})^2\Big)\end{aligned} \tag{5.7}$$

Note the negative sign - ϕ_c should be highest when the differences between parameters are smallest - and that each β is a vector so that every element of θ, τ and μ has its own coefficient . When an additional identity of y is given (i.e. "Given that someone is a teacher, what other identity is *that same person* most likely to also be?") the form of $\phi_y(i_y, i_{y'})$ is the same as $\phi_c(i_y, i_c)$, except that i_c is replaced with $i_{y'}$ and the weights are changed to $\beta_{\theta,y}$, $\beta_{\tau,y}$ and $\beta_{mu,y}$ respectively.

As argued in Section 5.3, defining a ϕ provides a definition of an identity labeling model; thus the expressed model here could in theory be used for identity labeling. However, the model as it is expressed is not useful in that we still have not defined what dimensions of the variables are important - that is, what institutional settings and what properties we should use for measurement.

A typical approach to this problem would be to simply assume there exists latent dimensions and then attempt to infer them. However, this approach is sub-optimal for two reasons. First, it does not utilize the significant domain knowledge that we have - for example, we know from decades of social science research that while we may not know what it is for a particular identity, the (assumed) gender of an identity is an important property of it. Second, and perhaps more basically, such a model would not be identifiable in a statistical sense.

Variable	Explanation
Q	The set of questions, q is a specific question
t_q	The type of question q ("IsA" or "SeenWith")
m_q	A selector variable for when an identity appears in the question text of q
S_q	The set of identities appearing as possible answers for q
z_q	A selector variable for when an identity is chosen as the answer for q
β_c	A vector incorporating $[\beta_{\theta,c}, \beta_{\tau,c}, \beta_{\mu,c}]$
β_y	The same as above, except for "IsA" questions
Θ	A vector containing the variables θ, τ, μ
K	A constant that defines a "score" associated with the "all are equally likely" option in the survey
$I()$	An indicator function

Table 5.10: Variables used in the section

We can address both of these problems by enforcing a limited amount of domain knowledge by defining axes of θ and τ based on *counter-identities*. For example, we could define a gender property in τ by introducing a dimension and ensuring that the identities "man" and "woman", a counter-identity pair, are at opposite ends of this dimension. This amounts to *fixing the value for counter-identity pairs* for different axes in the assumed latent space. As an example for association, we could define two axes for θ and fix the identity "doctor" at a particular value of one, defining the institution "Medicine", and fix the identity "sister" at a particular value of the other, defining the institution "Family" along that vector.

The intuition behind this is that we can define the space of properties and institutions by leveraging domain knowledge about the universe of identities available to us and defining them based on already known *relationships in meaning* between identity pairs. Recalling from above, this follows sociologically directly from Burke's (1980) work, and also fits nicely into the relational perspective developed in PCSM models. Intuitively, it also is a cognitive take on Blau's (1977) social space model, where we here represent each axes of our perceptions about identities by leveraging existing structure in the identity space. Finally, from a statistical perspective, fixing some set of values renders the model identifiable.

5.7.2 Estimating Parameters from Survey Data

One way to use the model we have developed is to infer semantic similarities and associations from the identity labeling questions in Study 2. In order to do so, we must be able to express answers to the survey as a function of the model. Using the variables in Table 5.10, we can do so by writing a likelihood function for the survey questions as follows:

$$p(Q|\Theta,\Sigma_c,\Sigma_y,\mathbf{m},\mathbf{S},\mathbf{z},K) =$$
$$\prod_q^Q \Bigg(\Big(\frac{K}{\sum_k^{S_q} \exp(\phi_c(m_q,k)) + K} \Big)^{I(t_q=SeenWith)} \Big(\frac{K}{\sum_k^{S_q} \exp(\phi_y(m_q,k)) + K} \Big)^{I(t_q=IsA)} \Big)^{I(z_q=none)}$$
$$\prod_i^I \Big(\frac{\exp(\phi_c(m_q,i))}{\sum_k^{S_q} \exp(\phi_c(m_q,k)) + K} \Big)^{I(t_q=SeenWith)} \Big(\frac{\exp(\phi_y(m_q,i))}{\sum_k^{S_q} \exp(\phi_y(m_q,k)) + K} \Big)^{I(t_q=IsA)} \Big)^{I(z_q=i)} \Bigg) \tag{5.8}$$

The likelihood is a product over all questions q in Q. For each question, the likelihood is divided into two parts. The first part represents the probability of the "all are equally (un)likely" option being selected by survey respondents - this occurs when $z_q = None$, in the equation, when $I(z_q = None) == 1$. The constant K, which we tune empirically, represents the weight (roughly, the probability) of this option relative to the respondent answering with one of the four identity labels. The second part is the probability model when an identity is selected as an answer. Both of these parts are also split into the different types of questions, for which we will use different weights.

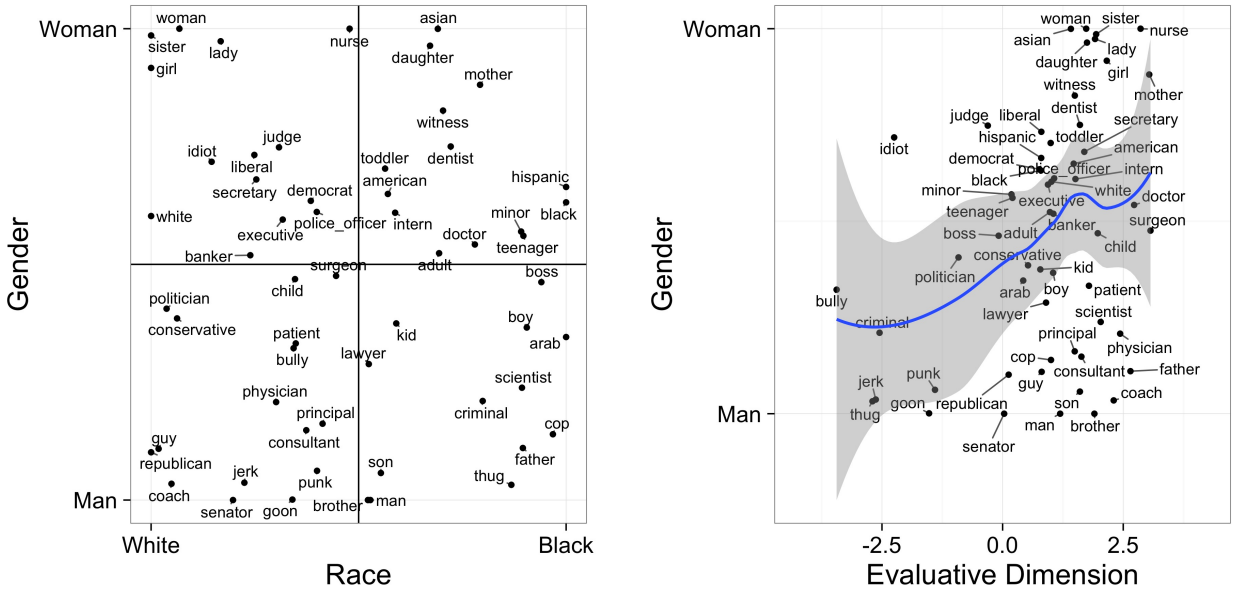
In order to maximize this likelihood function, we can "tune" three "knobs" - the underlying semantic and affective meanings of each identity - Θ- and the effect of semantic similarity, semantic association and affective similarity on the two different types of questions being asked - β_c and β_y. Appendix C presents the derivations necessary to do so.

By maximizing this likelihood function, we are able to understand how identities are similar and/or associated (via Θ), and also use β_c and β_y to reconsider the overall effects of semantic similarity, semantic association and affective similarity. All that remains is to define the parameters θ, τ; that is, to define what institutions and properties we are interested in and which identities we will fix the parameters of to make the model identifiable along these dimensions. We also will define μ by using the same affective dictionaries from the prior studies.

All columns of θ and τ are defined on a scale of -4.3 to 4.3 - note that as we discuss further in the appendix, optimization for θ is bounded within this range. For θ, we create six dimensions, one each for each institution in Table 5.7. We take one random identity from each institution and fix its value for a particular column at 4.3, effectively identifying that column as being representative of that identity.

For τ, we create six dimensions. On the first, we fix man to -4.3 and woman to 4.3, defining gender. On the second, we fix child to -4.3 and adult to 4.3, defining gender. On the third we fix white to -4.3 and black to 4.3, defining race. On the fourth we fix conservative to -4.3 and liberal to 4.3, defining political leaning. Finally, we create a two-dimensional race/ethnicity dimension, fixing Hispanic to the point (0, 4.3), Asian to the point (-2.48,0) and Arab to the point (2.48,0) - these points are equidistant in that space.

Finally, for the results shown below, we set $K = .1$. We also assume affective profiles are fixed, and thus do not estimate μ. We run the model on 90% of the data for 25 iterations of the optimization routine and select results from the iteration that minimized error on the 10% of the data we left out in order to avoid overfitting. Future work is necessary to develop a faster optimization routine that allows for a better mechanism of assessing convergence.



(a) (b)

Figure 5.10: a) presents estimates for the gender dimension of each identity label (y-axis) versus the race dimension of each identity label (x-axis). b) compares gender to the evaluative sentiment dimension of each identity.

5.7.3 Initial Results

Using our identity labeling model in this way essentially amounts to a factor analysis, decomposing identities into a latent space. Similar results could thus in theory be achieved by constructing an $|I|$x$|I|$ matrix of identities, where the value of each cell is the log-odds of selection for a particular type of question. The benefits of using the identity labeling model described are two-fold. First, results are based on an explicit theoretical model of identity labeling; parameters can thus be readily applied to other applications, such as predicting labeling in other contexts. Second, our model uses the available data in a better way than a factor analysis in the sense that we capture a direct likelihood equation for each question.

While several results are potentially of interest, we present two representative results of what I believe this model could be potentially useful for in the future. Figure 5.10a) presents each of the 57 identities modeled in two dimensions, one representing their estimated position on the race property and the other on the gender property. Results for identities like "guy", "sister", "girl", "mother", "brother", etc. confirm the models ability to infer intuitive alignments of identities to categories. More interestingly, the model also reveals significant implicit stereotypes played into the way that identity labeling occurred in the data from Study 2. In addition to the "secretary - woman" example we saw in Study 1, we see that "thugs" and "criminals" are seen as being more black than white, and that Republicans are being viewed as more male than female. These findings suggest that the model I have developed may be useful in uncovering how implicit stereotypes that go beyond denotive meanings may be driving the way we label other people.

Figure 5.10b) compares the evaluative dimension of an identity's profile to its gender. While

prior work has considered how gender relates to EPA profiles (Heise, 2007), here we are able to do so with not only obviously gendered identities but also subtly gendered identities. Doing so reveals a stronger relationship between gender and how good or bad an identity is perceived to be than if we would have considered only gendered identities - see, for example, the profiles of "son" and "father". The results in Figure 5.10b) suggest that the negative identities studied here are uniquely male-gendered; considering this result for a larger set of identities may be able to uncover the uniquely gendered negative identities for each gender as well.

5.8 Conclusion

In this chapter I consider the problem of how people label others and provide several contributions to this literature. First, I outline existing models of identity labeling, discussing their strengths and weaknesses. I then introduce a foundational mathematical model for identity labeling that is generalizes across the scope of existing literature and provides a jumping off point for future work on identity labeling.

Using this framework, I introduce an initial model of identity labeling that focuses on pairwise relationships between identities. I introduce how I expect semantic similarity, semantic association and affective similarity between identities to be important pillars of the identity labeling process. Importantly, I argue these relationships are relatively stable across different ways in which we are asked to label people, it is only how these quantities are used that changes. Thus, identity meanings and their relationships are consistent, measurable quantities.

Study 1 and Study 2 seem to confirm that semantic association and semantic similarity are important factors in how we label other people. Further, these studies suggest that we can decompose types of relationships between identities into four categories. Incomparable identities are not associated or similar - they are not identities we would ever think to combine in any way we are asked to label another individual or set of individuals. Synonymous identities refer to the same underlying concept - a person who holds one of the identities in the pair will by default hold the other. Counteridentity pairs represent identities that we tend to see together in social contexts but that have distinct properties/traits or occupy distinct ends of particular social categories. Finally, multiple identity pairs are identities that we would expect to apply to the same person but that we do not tend to apply to two individuals in a particular context. Interestingly, Study 2 suggested that most identities are incomparable - they make no sense together. Because of this, we agree with Heise (2007) that the scope of possible identities for any particular labeling task is small.

This finding may also help to confirm Heise and MacKinnon's (2010) view of a sort of two step process by which affective similarity is important in the identity labeling process, as well as helping to remedy the mixed results we see in this chapter with respect to affect. More specifically, I see results in this chapters as evidence for Heise and MacKinnon's (2010) assumption that semantic connections drive initial decisions on an identity's relevance to a situation, and affect is then more closely tied to behavioral decision making and action as a social situation ensues. The difference between the model we have postulated and what Heise and MacKinnon (2010), and Affect Control theorists in general, currently model is that we provide an explicit and cognitive model for this process. This result thus suggests a sequential model may be appropriate in linking the semantic model of Freeman and Ambady (2011) and the affective model of Heise

(2007).

Finally, I developed in this chapter a new mathematical model of identity labeling that goes beyond affective similarity, semantic similarity and semantic association to try and capture *how* identities are similar and associated (assuming we already know their affective profiles). Conceptually, the model unifies the idea of relational spaces of counteridentities from Burke (1980), ACT's assumptions of affective profiles of identities, and Freeman and Ambady's (2011) relational cognitive semantic model of identity labeling. Initial results of applying the model to survey data suggested that some work must be done to fit results better to the data, but that the model captures interesting elements of similarity, implicit stereotype and relations between certain axes of similarity and affective meaning that could not be studied otherwise.

Future work abounds from here. In the short term, addressing issues observed with estimating parameters of the proposed model from survey data should be addressed. I would also like to consider non-parametric forms of the model to ease the need for various assumptions made. Also, I would like to consider how predictive the model is in comparison to other possible models that could be proposed. Finally, I would like to better understand how I might be able to leverage NLP approaches and data to better initialize the model. In the longer term, I am interested in extending the proposed model to fully encapsulate behaviors, creating a stronger bond between the model that I propose and BayesACT. The model I develop also may be useful for cultural sociologists, as I see identity labeling as a relatively similar process to those considered in the study of culture. Finally, understanding how social interactions can impact perceptions as measured by this model is an exciting area of future work.

Chapter 6: Discussion

I have for much of my life been interested in how to reduce negative prejudices - that is, how to lessen the negative emotions individuals, groups, organizations and institutions carry of others. Prejudices are very closely tied to stereotypes[1], the meaning that we attribute to particular identities that individuals take on and a focal point of this thesis. An important downstream effect of stereotypes and prejudices are behaviors- how we feel about someone defines, to a large extent, how we will act towards them. Reducing negative prejudices is a complex problem - it concerns not only intergroup relationships, as it is often posed, but also has impacts on social issues like bullying, social isolation and mental health. Stereotypes and prejudice also exist in the human mind, which makes them notoriously hard to measure. Finally, to make matters worse, stereotypes are also encoded in and reaffirmed by the complex social structures within which we are ultimately embedded (Heise, 2007; Saperstein et al., 2013).

As recent work has suggested (Paluck, 2012), we need to better understand stereotypes and prejudice and how they exist within the human mind and within these sociocultural structures, before real progress on prejudice reduction can be made. This thesis introduces new theory, new methods and new datasets that serve this purpose. The work in this thesis gives us new ways to answer three critical questions in the study of identity and stereotype:

- What is the universe of identities available to a particular population?
- How do we measure the stereotypes attributed to these identities?
- How are these stereotypes used to apply identities to particular individuals in particular social situations?

Importantly, the methods and data developed are re-usable and, where possible, have been made public. This means that the tools presented can be utilized in their existing form, and that future researchers can adopt my approach and/or the data I have collected to create new methods and new theory. In concert with the development of these methods and the collection of this data, I have also developed a python package for the rapid collection and text processing of Twitter data. This tool, briefly described in Appendix A, has been used by several CASOS graduate students.

This thesis also makes contributions to the NLP literature by providing new datasets for method development, stronger justifications for existing approaches and a critical evaluation of existing datasets used currently for model evaluation. Finally, it presents some useful insight into two case studies interspersed throughout the document from newspaper and Twitter data on the Arab Spring and from Twitter data relevant to the Eric Garner and Michael Brown tragedies.

[1] They are, in my opinion, roughly equivalent based on how I have defined stereotypes quantitatively in this thesis, but I have chosen to postpone addressing this idiosyncracy and simply comply with the literature in this document

The contributions of this document consequently fall into both the computational domain and the domain of social psychology; or, if you prefer, into the evolving domain of computational social science. While these contributions are significant, this document ultimately opens up more doors for future work than it does close doors on existing questions.

One important avenue for follow-up work will be to better understand how results here generalize to different domains. For Chapters 3, 2 and 4, this means developing new models for different media or at least observing how these models apply in different domains. Unfortunately, the current models have been developed as end-to-end models for specific media, meaning that preprocessing, feature definitions and interpretations all must be adjusted to apply to different forms of text. A distant goal is also to expand beyond text data to consider similar questions of, for example, speech or video data. For Chapter 5, this means finding ways of moving beyond survey data and the two fairly specific situations that the survey questions were geared towards. An interesting way to do this would be to develop a game-like situation in which players could give labeling information as they play and game scenarios could be developed using active learning methods. This approach could also help to address limitations of the identity and stereotype extractions in Chapters 3, 2 and 4, making it a particularly interesting avenue for future work.

An equally important avenue for future work is integrating LCSS into models of the co-evolution of cognition and social structure (Carley, 1991) and the co-evolution of cognition, social structure, culture and institutions (Lizardo and Strand, 2010; Patterson, 2014). In order to make the theoretical model more useful, I have to answer questions about how the people we interact with impact our stereotypes and how our stereotypes impact who we interact with. Further, it is important to think about how our position in a social network impact our ability to enforce our biases on others or our willingness to change our biases. These issues will require new methods as well. I believe there is a strong link between these questions and the work being done on convolutional neural networks, but future work will be needed along these lines to confirm or deny this idea. A final note on this point - although I alluded to addressing some of these questions in my proposal, I quickly realized that this is a long-term goal that will require significantly more effort, time and money for survey data than I had at my disposal. The work that I did complete does serve, however, as an important jumping off point for this research, a point I address a bit further in Chapter 5.

Finally, the methodological work in this document could be extended to have stronger predictive power. One difficulty in this thesis was in developing models that had any predictive power at all that were still driven by social science and hence interpretable by social scientists. While the models I developed did beat the strong baselines that I pitted them against, there were many cases where the predictive abilities of my models were actually hindered by an adherence to social theory. On the one hand, this allowed me to make several interesting insights that may not have been possible using "black box" NLP methods. For exampe, Chapters 4 and 5 propose modifications to how we think about and measure semantic similarity; how well these observations can be used to inform the development of new neural embedding models is an important question. On the other hand, however, it is questionable how amenable these models in their current form will be to extension and adaptation by machine learning researchers interested purely in prediction. As this thesis has progressed, I have found literature that gives me more confidence that models which are both highly predictive and highly interpretable are on the horizon (e.g. McClelland, 2013), but for the moment these two goals are still somewhat at odds.

These three avenues for future work - extension to new media, combining LCSS with models of social networks and social institutions, and increasing the predictive power of the presented statistical models while retaining their interpretability for those outside the machine learning community - constitute the three biggest limitations I perceive of the present work. I also believe, however, that each is in and of itself an important undertaking deserving of its own multi-year research effort. As such, they are and the three most important next steps in my research agenda as I move forward into the next phase of my career. In addressing these questions, I hope to be more capable of addressing questions about how best to reduce negative prejudice and also to more closely tie the study of statistics and machine learning to the study of cognition and the social world.

Appendix A: Python Package `twitter_dm`

A.1 Brief Introduction

The python package `twitter_dm` is available at `https://github.com/kennyjoseph/twitter_dm`. The idea for the library was initially developed by Peter Landwehr but I have contributed almost all of the code and design of the package since late 2014. The library serves the following purposes:

- Rapidly collect data from the Twitter Search API. The `twitter_dm` library has the ability to leverage as many API connections, or "handles" to the API, as one has available to them. This is done via python's multiprocessing module, which we have extended to allow for a very flexible method to collect information on users and/or tweets. The `examples` directory in the package contains a bunch of examples of how this is done, and you can "extend" the classes in `twitter_dm/multiprocess` folder either by adding various function (e.g. to change how a snowball search is done) or by actually extending them in the traditional object-oriented programming sense. These classes are all subclasses of Twitter's multiprocessing library that help to do what we want to do fast and with a bunch of handles.
- Provide a repeatable way to tokenize, part-of-speech-tag, and dependency parse a Tweet in python. We use the great work by Brendan O'Connor, Myle Ott, Lingpeng Kong and others who did all the hard work of creating great Twitter NLP tools and put a few wrappers around them to hopefully make it slightly more easy to integrate into your workflow.
- Provide a convenient way to manipulate and access data from Twitter users and Tweets in an object-oriented, extendible fashion. The classes in `Tweet.py` and `TwitterUser.py` are convenient representations that make accessing data about users and tweets relatively painless.
- As a result of this thesis, the library also contains a way to extract social identities from text. The file `examples/run_identity_extractor.py` gives an end-to-end example of how to do this.

A.2 Getting Started

Once you have cloned the above repository, you can install the package and its dependencies by typing `python setup.py install` from the command line in the repository's directory.

Then, open up the `examples` folder and you will be provided with a number of examples on

ways to use the library to accomplish various tasks. I recommend starting with `collect_user_data_serial` for single-process collection of tweets, `snowball_sample_custom_function_example.py` for multithreaded collection of users with the ability to add a custom snow-ball sampling step and a custom data collection step and and `run_identity_extraction.py` for NLP stuff.

Anything where you're collecting from the API will require you to have access to the API - that is, you'll need a consumer key and secret (an application) and an access token and secret (a user who has subscribed to the API). `twitter_dm` can either help you set up new credentials (see `examples/auth_example.py`) or can handle credentials stored in text files, as described in the library's ReadMe file.

If one is interested in the object-oriented approach to Twitter data used by `twitter_dm`, you can also look at Tweet.py and TwitterUser.py.

Appendix B: Justification of Log-odds as an Outcome Variable

Recall that we only have survey data on similarity for the identity pairs in the SimLex-999 dataset. Consequently, we can only assume that similarity between the question identity and any other identity in our set comes from some distribution. Fortunately, this is enough to formally justify ignoring the value of the relationships between the question identity and any other identity in the question. Let us assume that the set of three randomly chosen identities for each question is given by the set S, and that the "all are equally (un)likely" has an expected constant probability that can therefore be ignored. Then, starting with Equation 5.2, we have:

$$p(i_y|y,c) = \frac{\phi(i_y,i_o)}{\phi(i_y,i_o)+\sum_{s\in S}\phi(i_s,i_o)} \approx \frac{\phi(i_y,i_o)}{\phi(i_y,i_o)+\sum_{s\in S}\mathbf{E}[\phi(i_s,i_o)]} \tag{B.1}$$

$$= \frac{\phi(i_y,i_o)}{\phi(i_y,i_o)+|S|\overline{\phi(i_s,i_o)}} = \frac{\phi(i_y,i_o)(1)}{\phi(i_y,i_o)(1+\frac{|S|\overline{\phi(i_s,i_o)}}{\phi(i_y,i_o)})} \tag{B.2}$$

$$= \frac{1}{1+|S|*\overline{\phi(i_s,i_o)}*\phi(i_y,i_o)^{-1}} \tag{B.3}$$

$$\propto \frac{1}{1+\phi(i_y,i_o)^{-1}} \tag{B.4}$$

The result of this proof, which relies on the assumption that S is a uniform random subset of I, is that the log-odds we calculate by ignoring all other answers will be proportional to the value we would get if we had all information about all pairs of identities. In other words, we can safely assume that the ordering of log-odds that we see would be no different, in expectation, than if we were to incorporate in relationships with the randomized set of identities per question.

Note that we are, as one might expect, not the first to observe this connection. In particular, our observation is directly related to the notion of *case-controlled* approximate likelihood calculations in epidemiology (Breslow, 1996)

Appendix C: Maximum Likelihood Estimation of the Model for Study 2

Let us first express the loglikelihood of the model:

$$\ell(\Theta,\beta) = \sum_{q}^{Q} I(z_q = none)\Big(\log(K) - (I(t_q = SeenWith) * \log(\sum_{k}^{S_q}\exp(\phi_c(m_q,k)) + K) + I(t_q = IsA) * \log(\sum_{k}^{S_q}\exp(\phi_c(m_q,k)) + K))\Big) + \sum_{i}^{I} I(z_q = i)\Bigg(I(t_q = SeenWith) * \Big(\phi_c(m_q,i) - \log(\sum_{k}^{S_q}\exp(\phi_c(m_q,k)) + K)\Big) + I(t_q = IsA)\Big((\phi_y(m_q,i) - \log(\sum_{k}^{S_q}\exp(\phi_y(m_q,k)) + K)\Big)\Bigg) \tag{C.1}$$

We are unable to analytically solve for the maximum likelihood estimate. Instead, we will use coordinate descent on β_c, β_y and Θ to come to a maximum likelihood estimate. To do so, we will use a bounded quasi-Newton method, L-BFGS-B, to optimize this function, restricting the domain of Θ to (-4.3,4.3). We will iteratively optimize over β_c, β_y and each row of Θ (i.e. the parameters for each identity independently) until convergence. In order to do so efficiently we derive the gradient for each parameter; we allowed the Hessian to be determined numerically in the optimization tool we used. Let us consider the gradient for the weights for "SeenWith" questions, β_c - derivations for β_y are analogous:

$$\frac{\delta\ell}{\delta\beta_c} = \sum_{q}^{Q} I(z_q = none)I(t_q = SeenWith)\Big(\frac{\sum_k^{S_k}(\Theta_{m_q} - \Theta_k)^2\exp(-\beta_c'(\Theta_{m_q} - \Theta_k)^2)}{\sum_k^{S_k}\exp(-\beta_c'(\Theta_{m_q} - \Theta_k)^2) + K}\Big) + \sum_{i}^{I} I(z_q = i)I(t_q = SeenWith)\Big(-(\Theta_{m_q} - \Theta_i)^2 + \frac{\sum_k^{S_k}(\Theta_{m_q} - \Theta_k)^2\exp(-\beta_c'(\Theta_{m_q} - \Theta_k)^2)}{\sum_k^{S_k}\exp(-\beta_c'(\Theta_{m_q} - \Theta_k)^2) + K}\Big) \tag{C.2}$$

We now consider the gradient for each identity's row in Θ. Let us consider Θ for a particular

identity j, thus taking a single row in the matrix Θ, and assume we are focused only on "Seen-With" questions - results for "IsA" questions are analogous. We must consider four different cases. In the first case, j is not in the text of a question q (i.e. $j \neq m_q$) nor is j given as one of the multiple choice questions (i.e. $j \notin S_q$). We can simply ignore these questions. We represent the other three cases below, where taking the derivative with respect to Θ_j results in different values. Let $\ell(q) = \phi_c(m_q, z_q) - \log(\sum_k^{S_q} \exp(\phi_c(m_q, k))$.

If $j = m_q$, then:

$$\frac{\delta \ell(q)}{\delta \Theta_j} = -2\beta_c'(\Theta_j - \Theta_{z_q}) + \frac{\sum_k^{Y_k} 2\beta_c'(\Theta_j - \Theta_k) \exp(-\beta_c'(\Theta_j - \Theta_k)^2)}{\sum_k^{Y_k} \exp(-\beta_c'(\Theta_j - \Theta_k)^2)} \tag{C.3}$$

If $j = z_q$, then:

$$\begin{aligned} \frac{\delta \ell(q)}{\delta \Theta_j} &= 2\beta_c'(\Theta_{m_q} - \Theta_j) - \frac{2\beta_c'(\Theta_{m_q} - \Theta_j) \exp(-\beta_c'(\Theta_{m_q} - \Theta_j)^2)}{\sum_k^{Y_k} \exp(-\beta_c'(\Theta_{m_q} - \Theta_k)^2)} \\ &= 2\beta_c'(\Theta_{m_q} - \Theta_j)\Big(1 - \frac{\exp(-\beta_c'(\Theta_{m_q} - \Theta_j)^2)}{\sum_k^{Y_k} \exp(-\beta_c'(\Theta_{m_q} - \Theta_k)^2)}\Big) \end{aligned} \tag{C.4}$$

Finally, if $j \neq z_q$ but $j \in S_q$, then:

$$\frac{\delta \ell(q)}{\delta \Theta_j} = -\frac{2\beta_c'(\Theta_{m_q} - \Theta_j) \exp(-\beta_c'(\Theta_{m_q} - \Theta_j)^2)}{\sum_k^{Y_k} \exp(-\beta_c'(\Theta_{m_q} - \Theta_k)^2)} \tag{C.5}$$

If the answer is none, the derivation for all but the second case is the exact same (the answer can't be none in the third base); we just don't add the front term.

Using the results of Equation C.3, C.4 and C.5, computing the gradient for Θ_j is a straightforward addition over these various cases across both types of cases - the full equation is not shown here because of this.

Bibliography

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paca, M., and Soroa, A. (2009). A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics. 5.2.3

Ahmed, A., Hong, L., and Smola, A. J. (2013). Hierarchical geographical modeling of user locations from social media posts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 25–36. 3.2.2

Ahothali, A. and Hoey, J. (2015). Good News or Bad News: Using Affect Control Theory to Analyze Readers Reaction Towards News Articles. ACL. 2.2.2, 3.1, 4, 4.1.2, 5.4.1

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of verbal learning and verbal behavior*, 22(3):261–295. 5.2.3

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford University Press, Oxford [etc.]. 3.6.2, 5.2.3

Bamman, D. (2015). *People-Centric Natural Language Processing*. PhD thesis, Carnegie Mellon University. 3.2.2

Bamman, D. and Smith, N. A. (2014). Unsupervised Discovery of Biographical Structure from Text. *TACL*. 4.1

Bamman, D., Underwood, T., and Smith, N. A. (2014). A bayesian mixed effects model of literary character. In *Proceedings of the 52st Annual Meeting of the Association for Computational Linguistics (ACL14)*. 1, 3.1, 4.1, 4.1.1

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516. 4.4.2

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Dont count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 238–247. 4.1.1

Bertrand, M. and Mullainathan, S. (2003). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. Technical report, National Bureau of Economic Research. 1, 4

Beukeboom, C. J. and others (2014). Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies. *Social cognition and communication*, 31:313–330. 4.1

Bishop, C. M. and others (2006). *Pattern recognition and machine learning*, volume 1. springer New York. 9

Blau, P. (1977). A macrosociological theory of social structure. *American journal of sociology*, pages 26–54. 5.2.1, 5.5.4, 5.7.1

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. 5.2.3

Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35. 1, 4.1.1, 4.4.2, 5.2.3

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022. 2.2.2, 2.8, 3.1, 3.6.2, 5.2.3

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008. 4.4.2

Bosagh Zadeh, R., Goel, A., Munagala, K., and Sharma, A. (2013). On the precision of social and information networks. In *Proceedings of the first ACM conference on Online social networks*, pages 63–74. 3.6.1, 3.6.1

Bradley, J. R. (2012). *After the Arab Spring: How Islamists Hijacked The Middle East Revolts*. Macmillan. 2.7.2

Breslow, N. E. (1996). Statistics in epidemiology: the case-control study. *Journal of the American Statistical Association*, 91(433):14–28. B

Bromiley, P. A. (2013). *Products and Convolutions of Gaussian Probability Density Functions*. Tina-Vision Memo. 10

Bucholtz, M. and Hall, K. (2005). Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, 7(4-5):585–614. 5.1

Burke, P. J. (1980). The self: Measurement requirements from an interactionist perspective. *Social psychology quarterly*, pages 18–29. 1, 5.1, 5.2.1, 5.2.3, 5.3, 5.7.1, 5.7.1, 5.7.1, 5.8

Butts, C. T. (2008). A relational event framework for social action. *Sociological Methodology*, 38(1):155–200. 2.8

Cammett, M. and Luong, P. J. (2014). Is there an islamist political advantage? *Annual Review of Political Science*, 17:187–206. 2.7.2

Carley, K. (1991). A Theory of Group Stability. *American Sociological Review*, 56(3):331–354. ArticleType: research-article / Full publication date: Jun., 1991 / Copyright 1991 American Sociological Association. 6

Carley, K. (1994). Extracting culture through textual analysis. *Poetics*, 22(4):291–312. 2.8

Carreras, X. and Màrquez, L. (2005). Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 152–164. Association for Computational Linguistics. 2.3

Chambers, N., Bowen, V., Genco, E., Tian, X., Young, E., Harihara, G., and Yang, E. (2013). Identifying Political Sentiment between Nation States with Social Media. *Computational Linguistics*, 39:4. 4.1.2

Charles, W. G. (2000). Contextual correlates of meaning. *Applied Psycholinguistics*, 21(04):505–

524. 5.2.3

Charniak, E. (1996). *Statistical language learning*. MIT press. 2.4

Chen, J., Zhu, J., Wang, Z., Zheng, X., and Zhang, B. (2013). Scalable inference for logistic-normal topic models. In *Advances in Neural Information Processing Systems*, pages 2445–2453. 4.3.2, 4.3.3, 4.3.4

Chen, L., Wang, W., Nagarajan, M., Wang, S., and Sheth, A. P. (2012). Extracting Diverse Sentiment Expressions with Target-Dependent Polarity from Twitter. In *ICWSM*. 1, 4, 4.1.2

Cikara, M. and Van Bavel, J. J. (2014). The Neuroscience of Intergroup Relations An Integrative Review. *Perspectives on Psychological Science*, 9(3):245–274. 1, 3.2.1, 5.1

Collins, A. M. and Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407. 5.2.2, 5.2.3

Cuddy, A. J., Fiske, S. T., and Glick, P. (2007). The BIAS map: behaviors from intergroup affect and stereotypes. *Journal of personality and social psychology*, 92(4):631. 5.2.1

Davenport, L. D. (2016). The Role of Gender, Class, and Religion in Biracial Americans Racial Labeling Decisions. *American Journal of Sociology*, 81(1):57–84. 4.2, 5.1, 5.2.2, 5.4.2

De Marneffe, M.-C. and Manning, C. D. (2008). Stanford typed dependencies manual. *URL http://nlp. stanford. edu/software/dependencies manual. pdf*. 2.3

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS*, 41(6):391–407. 2.2.2, 5.2.3

Deese, J. (1966). *The structure of associations in language and thought*. Johns Hopkins University Press. 4.1.1

Del Corro, L., Abujabal, A., Gemulla, R., and Weikum, G. (2015). FINET: Context-Aware Fine-Grained Named Entity Typing. *EMNLP'15*. 3.1, 3.2.2, 3.4.2, 4.1.1

DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data & Society*, 2(2):2053951715602908. 4.3.3

Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M., and Xu, K. (2014). Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 49–54. 4.1.2

Dovidio, J. F. and Gaertner, S. L. (1999). Reducing Prejudice Combating Intergroup Biases. *Current Directions in Psychological Science*, 8(4):101–105. 5.2.1

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1277–1287, Stroudsburg, PA, USA. Association for Computational Linguistics. 3.2.2

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2014). Diffusion of Lexical Change in Social Media. *PloS one*, 9(11):e113114. 4.3

Ekbal, A., Sourjikova, E., Frank, A., and Ponzetto, S. P. (2010). Assessing the Challenge of Fine-grained Named Entity Recognition and Classification. In *Proceedings of the 2010 Named Entities Workshop*, NEWS '10, pages 93–101, Stroudsburg, PA, USA. Association for Computational Linguistics. 3.1, 3.2.2, 4.1.1

Fast, E., Vachovsky, T., and Bernstein, M. S. (2016). Shirtless and Dangerous: Quantifying

Linguistic Signals of Gender Bias in an Online Fiction Writing Community. *arXiv preprint arXiv:1603.08832*. 4.1

Fiske, S. T., Cuddy, A. J., Glick, P., and Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of personality and social psychology*, 82(6):878. 1, 4, 5.2.1

Flaxman, S. R., Neill, D. B., and Smola, A. J. (2015). Gaussian Processes for Independence Tests with Non-iid Data in Causal Inference. *Provisional acceptance at ACM Transactions on Intelligent Systems and Technology (TIST), 2015b. URL http://www. sethrf. com/files/gp-depend. pdf*. 5.5.3

Fleischman, M. and Hovy, E. (2002). Fine grained classification of named entities. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics. 3.1, 3.3.2

Fraley, C., Raftery, A. E., Murphy, T. B., and Scrucca, L. (2012). mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. Technical report, Technical Report. 17

Francis, C. and Heise, D. R. (2006). Mean affective ratings of 1,500 concepts by indiana university undergraduates in 2002–3. *Data in Computer Program Interact*. 1, 3, 2.5.4

Freeman, J. B. and Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological review*, 118(2):247. 1, 4, 4.1.1, 4.1.3, 5.1, 5.2.2, 5.3, 5.7.1, 5.8

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441. 4.4.2

Gabrilovich, E. and Markovitch, S. (2007). Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *IJCAI*, volume 7, pages 1606–1611. 5.2.3

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*, volume Third Edition. CRC press. 9, 2.4.2

Goffman, E. (1959). *The Presentation of Self in Everyday Life*. Anchor. 3.1

Graells-Garrido, E., Lalmas, M., and Menczer, F. (2015). First Women, Second Sex: Gender Bias in Wikipedia. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 165–174. ACM. 4.1

Greenwald, A. G. and Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102(1):4–27. 4.5, 5.2.2, 5.7.1

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proc. of the National Academy of Sciences*, 101(suppl 1):5228–5235. 4.3.3

Grycner, A., Weikum, G., Pujara, J., Foulds, J., and Getoor, L. (2015). RELLY: Inferring Hypernym Relationships Between Relational Phrases. 3.5.2

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, volume 43. CRC Press. 5.5.3

Heise, D. R. (1979). *Understanding events: Affect and the construction of social action*. CUP Archive. 2.2.1, 5.1

Heise, D. R. (1987). Affect control theory: Concepts and model. *The Journal of Mathematical Sociology*, 13(1-2):1–33. 1, 2.1, 2.2.1, 5.1

Heise, D. R. (2007). *Expressive Order*. Springer. 1, 2.1, 2.2.1, 2.2.1, 2.5.2, 2.5.4, 2.8, 3.1, 3.3.1, 4, 4.1.2, 4.1.3, 4.3.3, 4.4.2, 5.1, 5.2.1, 5.4.1, 5.6, 5.7.1, 5.7.3, 5.8, 6

Heise, D. R. (2010a). INTERACT: Introduction and software. 1, 2.5.4, 3.1, 3.3.2, 3.6.2

Heise, D. R. (2010b). *Surveying cultures: Discovering shared conceptions and sentiments*. John Wiley & Sons. 4

Heise, D. R. (2014). Determinants of normative processes: comparison of two empirical methods of specification. *Quality & Quantity*, pages 1–18. 2.2.1, 2.2.1

Heise, D. R. and MacKinnon, N. J. (2010). *Self, identity, and social institutions*. Palgrave Macmillan. 1, 2.8, 3.2.1, 3.7.2, 3.8, 4, 4.1.1, 5.1, 5.2.1, 5.2.3, 5.6, 5.7.1, 5.8

Hill, F., Cho, K., Jean, S., Devin, C., and Bengio, Y. (2014a). Not All Neural Embeddings are Born Equal. *arXiv preprint arXiv:1410.0718*. 5.2.3, 5.5.3, 5.5.3, 5.5.4, 5.6, 5.6.2

Hill, F., Reichart, R., and Korhonen, A. (2014b). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *arXiv preprint arXiv:1408.3456*. 4.1.1, 1, 5.4.1

Hoang, T.-A., Cohen, W. W., and Lim, E.-P. (2014). On modeling community behaviors and sentiments in microblogging. SIAM. 3.2.2

Hoey, J. and Schröder, T. (2015). Bayesian affect control theory of self. In *Proc. of the AAAI Conference on Artificial Intelligence*. 5.2.1

Hoey, J., Schröder, T., and Alhothali, A. (2013a). Affect control processes: Intelligent affective interaction using a partially observable markov decision process. *arXiv:1306.5279 [cs]*. arXiv: 1306.5279. 1, 2.1, 2.2.1, 2.6

Hoey, J., Schroder, T., and Alhothali, A. (2013b). Bayesian Affect Control Theory. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 166–172. 2.2.1, 4.1.2, 3, 5.2.1, 5.3

Hoffart, J., Suchanek, F. M., Berberich, K., Lewis-Kelham, E., De Melo, G., and Weikum, G. (2011). YAGO2: exploring and querying world knowledge in time, space, context, and many languages. In *Proceedings of the 20th international conference companion on World wide web*, pages 229–232. ACM. 5.2.3

Hoffart, J., Suchanek, F. M., Berberich, K., and Weikum, G. (2013). YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, 194:28–61. 5.2.3

Hovy, D., Zhang, C., Hovy, E., and Peas, A. (2011). Unsupervised Discovery of Domain-specific Knowledge from Text. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1466–1475, Stroudsburg, PA, USA. Association for Computational Linguistics. 3.1, 3.2.2

Hutto, C. J. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAAI Conference on Weblogs and Social Media*. 3.3.1, 3.6.2, 4.1.2, 4.3.3

Jiang, L., Yu, M., Zhou, M., Liu, X., and Zhao, T. (2011). Target-dependent twitter sentiment classification. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 151–160. Association for Computational Linguistics. 4.1.2

Johnson, K. L., Freeman, J. B., and Pauker, K. (2012). Race is gendered: how covarying pheno-

types and stereotypes bias sex categorization. *Journal of personality and social psychology*, 102(1):116. 5.1

Joseph, K., Carley, K. M., Filonuk, D., Morgan, G. P., and Pfeffer, J. (2014). Arab Spring: from newspaper data to forecasting. *Social Network Analysis and Mining*, 4(1):1–17. 2.3

Joseph, K., Wei, W., Benigni, M., and Carley, K. M. (2016a). Inferring affective meaning from text using Affect Control Theory and a probabilistic graphical model. *Journal of Mathematical Sociology*. 1, 3.1, 3, 4.3.2, 4.3.3, 5.4.2

Joseph, K., Wei, W., and Carley, K. M. Girls rule, boys drool: Inferring semantic and affective relationships between identities from text. In *ICWSM*. 5.4.1

Joseph, K., Wei, W., and Carley, K. M. (2016b). Exploring patterns of identity usage in tweets: a new problem, solution and case study. In *WWW*. 1, 4, 4.1.1, 4.1.2, 4.2, 4.3.1, 5.2.3, 5.4.2

Kang, S. K. and Bodenhausen, G. V. (2015). Multiple Identities in Social Perception and Interaction: Challenges and Opportunities. *Annual Review of Psychology*, 66(1):547–574. 5.1

Katz, D. and Braly, K. (1933). Racial stereotypes of one hundred college students. *The Journal of Abnormal and Social Psychology*, 28(3):280–290. 1

Kiritchenko, S., Zhu, X., and Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, pages 723–762. 4.1.2

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press. 2.1

Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., and Smith, N. A. (2014). A dependency parser for tweets. In *Proc. of the Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, to appear*. 3.1, 3.4.3, 4.3.1

Kralj Novak, P., Smailovic, J., Sluban, B., and Mozetic, I. (2015). Sentiment of Emojis. *arXiv preprint arXiv:1509.07761*. 4.2

Kübler, S., McDonald, R., and Nivre, J. (2009). Dependency parsing. *Synthesis Lectures on Human Language Technologies*, 1(1):1–127. 2.1, 2.3

Kunda, Z. and Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review*, 103(2):284–308. 1, 4, 4.1.1, 5.1, 5.2.2

Le, M. N. and Fokkens, A. (2016). Taxonomy Beats Corpus in Similarity Identification, but Does It Matter? In *Recent advances in NLP*, page 346. 5.2.3

Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 609–616. ACM. 2.2.2

Leetaru, K. and Schrodt, P. A. (2013). Gdelt: Global data on events, location, and tone, 1979–2012. In *of: Paper presented at the ISA Annual Convention*, volume 2, page 4. 2.1

Levy, O. and Goldberg, Y. (2014). Dependencybased word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 302–308. 4.1.1, 5.2.3, 5.6.2

Li, W. and McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, pages

577–584. 5.2.3

Lin, T., Etzioni, O., and others (2012). No noun phrase left behind: detecting and typing unlinkable entities. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 893–903. Association for Computational Linguistics. 3.1

Liu, K.-L., Li, W.-J., and Guo, M. (2012). Emoticon smoothed language models for twitter sentiment analysis. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*. 4.1.2

Liu, Q., Jiang, H., Wei, S., Ling, Z.-H., and Hu, Y. (2015). Learning semantic word embeddings based on ordinal knowledge constraints. *Proceedings of ACL, Beijing, China*. 4.5

Lizardo, O. and Strand, M. (2010). Skills, toolkits, contexts and institutions: Clarifying the relationship between different approaches to cognition in cultural sociology. *Poetics*, 38(2):205–228. 5.1, 5.2.1, 6

Lui, M. and Baldwin, T. (2012). langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30. Association for Computational Linguistics. 3

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 142–150. Association for Computational Linguistics. 1, 4.1.3, 4.5

Malik, M. M., Lamba, H., Nakos, C., and Pfeffer, J. (2015). Population Bias in Geotagged Tweets. In *Ninth International AAAI Conference on Web and Social Media*. 4.5

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60. 2.3

Marwick, A. E. and Boyd, D. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, 13(1):114–133. 4.4.2

McCallum, A., Wang, X., and Corrada-Emmanuel, A. (2007). Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, 30(1):249–272. 5.2.3

McClelland, J. L. (2013). Integrating probabilistic models of perception and interactive neural networks: a historical and tutorial review. 6

McFadden, D. (1980). Econometric models for probabilistic choice among products. *Journal of Business*, pages 13–29. 5.3

Medelyan, O., Witten, I. H., Divoli, A., and Broekstra, J. (2013). Automatic construction of lexicons, taxonomies, ontologies, and other knowledge structures. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(4):257–279. 5.2.3

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. 4.1.1, 5.2.3, 5.6.2

Mikolov, T., Yih, W.-t., and Zweig, G. (2013b). Linguistic Regularities in Continuous Space Word Representations. In *HLT-NAACL*, pages 746–751. Citeseer. 2.2.2, 3.3.3

Miles, A. (2014). Addressing the Problem of Cultural Anchoring An Identity-Based Model of Culture in Action. *Social Psychology Quarterly*, 77(2):210–227. 5.1

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41. 3.2.1, 5.2.3

Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28. 4.1.1, 4.5, 5.2.3

Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Betteridge, J., Carlson, A., Dalvi, B., Gardner, M., Kisiel, B., Krishnamurthy, J., Lao, N., Mazaitis, K., Mohamed, T., Nakashole, N., Platanios, E., Ritter, A., Samadi, M., Settles, B., Wang, R., Wijaya, D., Gupta, A., Chen, X., Saparov, A., Greaves, M., and Welling, J. (2015). Never-Ending Learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*. 5.2.3

Mohammad, S. M. and Turney, P. D. (2010). Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics. 4.2

Mohr, J. W. and Bogdanov, P. (2013). IntroductionTopic models: What they are and why they matter. *Poetics*, 41(6):545–569. 5

Morgan, J. H., Rogers, K. B., and Hu, M. (2015). Distinguishing Normative Processes from Noise: A Comparison of Four Approaches to Modeling Impressions of Social Events. In Submission. 4.1.2

Neal, R. M. (2003). Slice sampling. *Annals of statistics*, pages 705–741. 2.8

Nelson, D. L., McEvoy, C. L., and Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407. 5.2.3, 5.4.1

O'Connor, B., Stewart, B. M., and Smith, N. A. (2013). Learning to Extract International Relations from Political Context. In *ACL (1)*, pages 1094–1104. 2.1, 2.2.2, 2.3, 4.3.3

Osgood, C. E. (1969). On the whys and wherefores of E, P, and A. *Journal of Personality and Social Psychology*, 12(3):194. 5.4.1

Osgood, C. E. (1975). *Cross-cultural universals of affective meaning*. University of Illinois Press. 2.1, 2.2.1

Owens, T. J., Robinson, D. T., and Smith-Lovin, L. (2010). Three faces of identity. *Sociology*, 36(1):477. 5.1, 5.2.1

Owoputi, O., OConnor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL*. 3.3.1, 3.4.3, **??**, 5.4.2

Paluck, E. L. (2012). Interventions Aimed at the Reduction of Prejudice and Conict. In *The Oxford Handbook of Intergroup Conflict*, pages 179–192. Oxford University Press. 6

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135. 2.1, 2.2.2

Patterson, O. (2014). Making Sense of Culture. *Annual Review of Sociology*, (0). 6

Penner, A. M. and Saperstein, A. (2013). Engendering Racial Perceptions An Intersectional

Analysis of How Social Status Shapes Race. *Gender & Society*, 27(3):319–344. 5.1, 5.2.2

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12:1532–1543. 3.1, 3.3.3, 10, 4, 4.1.1, 4.4.2, 5.6.2

Pettigrew, T. F. and Tropp, L. R. (2008). How does intergroup contact reduce prejudice? Meta-analytic tests of three mediators. *European Journal of Social Psychology*, 38(6):922–934. 3.6.2

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 5.5.3

Rajaraman, A. and Ullman, J. D. (2011). *Mining of massive datasets*. Cambridge University Press. 2.3

Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 248–256. Association for Computational Linguistics. 3.2.2

Read, S. J. and Miller, L. C. (1998). On the dynamic construction of meaning: An interactive activation and competition model of social perception. *Connectionist models of social reasoning and social behavior*, pages 27–68. 1, 5.1

Recasens, M., Hovy, E., and Mart, M. A. (2011). Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138–1152. 3.2.1, 5.1

Rehurek, R. and Sojka, P. (2011). GensimPython Framework for Vector Space Mo delling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*. 3.6.2

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.(JAIR)*, 11:95–130. 4.1.1, 5.2.2, 5.2.3, 3

Ritter, A., Clark, S., and Etzioni, O. (2011). Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics. 3.1, 3.2.2, 3.4.2, 3.4.3, 3.5.2

Roberts, M. E., Stewart, B. M., and Airoldi, E. M. (2013). Structural topic models. Technical report, Working paper. 2.8

Robinson, D. T., Smith-Lovin, L., and Wisecup, A. K. (2006). Affect Control Theory. In *Handbook of the Sociology of Emotions*, Handbooks of Sociology and Social Research, pages 179–202. Springer US. 1, 2.1, 2.2.1, 4.1.3, 5.1

Rogers, K. B., Schrder, T., and Scholl, W. (2013). The Affective Structure of Stereotype Content Behavior and Emotion in Intergroup Context. *Social Psychology Quarterly*, 76(2):125–150. 1, 4.1.2, 4.4.2, 5.2.1

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S. M., Ritter, A., and Stoyanov, V. (2015). Semeval-2015 task 10: Sentiment analysis in twitter. In *Proc. of the 9th International Workshop on Semantic Evaluation, SemEval*. 4.1.2

Ruths, D. and Pfeffer, J. (2014). Social media for large studies of behavior. *Science*,

346(6213):1063–1064. 4.5

Sahlgren, M. (2006). The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. 4.1.1, 1, 5.2.3

Saperstein, A., Penner, A. M., and Light, R. (2013). Racial Formation in Perspective: Connecting Individuals, Institutions, and Power Relations. *Annual Review of Sociology*, 39(1):359–378. 3.2.1, 6

Schröder, T., Hoey, J., and Rogers, K. B. (2017). Modeling dynamic identities and uncertainty in social interactions: Bayesian affect control theory. *Am. Soc. Rev*. 5.1, 5.2.1

Schröder, T., Rogers, K. B., Ike, S., Mell, J. N., and Scholl, W. (2013). Affective meanings of stereotyped social groups in cross-cultural comparison. *Group processes & intergroup relations*, page 1368430213491788. 5.2.2, 5.4.1, 5.4.1, 5.4.2

Schröder, T., Stewart, T. C., and Thagard, P. (2014). Intention, emotion, and action: A neural theory based on semantic pointers. *Cognitive science*, 38(5):851–880. 5.1, 5.2.1, 5.2.2

Schröder, T. and Thagard, P. (2014). Priming: Constraint satisfaction and interactive competition. *Understanding Priming Effects in Social Psychology*, page 157. 1, 4.1.1, 5.1

Shalizi, C. R. (2013). Advanced data analysis from an elementary point of view. *URL: http://www. stat. cmu. edu/cshalizi/ADAfaEPoV/13*, 24. 16

Smith-Lovin, L. (1987a). The affective control of events within settings*. *Journal of Mathematical Sociology*, 13(1-2):71–101. 5.3

Smith-Lovin, L. (1987b). Impressions from events. *Journal of Mathematical Sociology*, 13(1-2):35–70. 2.2.1

Smith-Lovin, L. (2007). The Strength of Weak Identities: Social Structural Sources of Self, Situation and Emotional Experience. *Social Psychology Quarterly*, 70(2):106–124. 1, 3.2.1, 4, 5.1

Smith-Lovin, L. and Douglas, W. (1992). An affect control analysis of two religious subcultures. *Social perspectives on emotion*, 1:217–47. 2.2.1, 3.7.2, 5.2.1, 5.4.1

Smith-Lovin, L. and Robinson, D. T. (2015). Interpreting and Responding to Events in Arabic Culture. *Final Report to Office of Naval Research, Grant N00014-09-1-0556*. (document), 4.2, 5.4.1, 5.1, 5.4.1, 5.4.2

Soon, W. M., Ng, H. T., and Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544. 2.3

Stenetorp, P., Pyysalo, S., Topic, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics. 3.4.1

Stets, J. E. and Carter, M. J. (2012). A Theory of the Self for the Sociology of Morality. *American Sociological Review*, 77(1):120–140. 5.2.1

Stryker, S. and Burke, P. J. (2000). The past, present, and future of an identity theory. *Social psychology quarterly*, pages 284–297. 5.1, 5.2.1

Taddy, M. (2013). Measuring Political Sentiment on Twitter: Factor Optimal Design for Multi-

nomial Inverse Regression. *Technometrics*, 55(4):415–425. 4.1.2

Tajfel, H. and Turner, J. C. (1979). An integrative theory of intergroup conflict. In *The social psychology of intergroup relations*, pages 33–47. Brooks/Cole, Monterey, CA, w austin & s. worche edition. 2.1, 3.6.2, 5.1, 5.2.1

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., and Qin, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1555–1565. 4.1.3, 4.5

Teh, Y. W. (2010). Dirichlet process. In *Encyclopedia of machine learning*, pages 280–287. Springer. 2.8

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2012). Hierarchical dirichlet processes. *Journal of the american statistical association*. 5.2.3

Thomas, L. and Heise, D. R. (1995). Mining Error Variance and hitting pay-dirt: Discovering systematic variation in social sentiments. *The Sociological Quarterly*, 36(2):425–439. 2.2.1

Tregouet, D. A., Escolano, S., Tiret, L., Mallet, A., and Golmard, J. L. (2004). A new algorithm for haplotype-based association analysis: the stochastic-EM algorithm. *Annals of human genetics*, 68(2):165–177. 9

Tufekci, Z. (2014). Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. In *ICWSM 14: Proc. of the 8th International AAAI Conference on Weblogs and Social Media.* 3.3.1

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., and Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory.* Basil Blackwell, Cambridge, MA. 3.6.2, 5.2.1

Tversky, A. and Gati, I. (1978). Studies of similarity. *Cognition and categorization*, 1(1978):79–98. 5.1, 5.2.2, 5.2.3

Vaisey, S. (2014). Is interviewing compatible with the dual-process model of culture. *American Journal of Cultural Sociology*, 2(1):150–158. 5.5.4

Van Bavel, J. J. and Cunningham, W. A. (2010). A social neuroscience approach to self and social categorisation: A new look at an old issue. *European Review of Social Psychology*, 21(1):237–284. 1, 5.1

Vo, D.-T. and Zhang, Y. (2015). Target-dependent twitter sentiment classification with rich automatic features. In *Proc. of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pages 1347–1353. 4.1.3

Vrandei, D. and Krtzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85. 5.2.3

Wallach, H., Mimno, D., and McCallum, A. (2009). Rethinking LDA: Why priors matter. *Advances in Neural Information Processing Systems*, 22:1973–1981. 2.5.4

Wang, P., Robins, G., Pattison, P., and Lazega, E. (2013). Exponential random graph models for multilevel networks. *Social Networks*, 35(1):96–115. 4.4.2

Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, 45(4):1191–1207. (document), 4.2, 5.4.1, 5.4.1, 5.1

Wei, W., Joseph, K., Liu, H., and Carley, K. M. (2015a). The Fragility of Twitter Social Networks Against Suspended Users. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 9–16. ACM. 3.7.1

Wei, W., Joseph, K., Lo, W., and Carley, K. M. (2015b). A Bayesian Graphical Model to Discover Latent Events from Twitter. In *Ninth International AAAI Conference on Web and Social Media*. 3.2.2

Wood, S. (2006). *Generalized additive models: an introduction with R*. CRC press. 5.5.3

Wu, Z. and Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics. 5.6.2

Yang, L., Sun, T., Zhang, M., and Mei, Q. (2012). We know what @you #tag: does the dual role affect hashtag adoption? In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 261–270, New York, NY, USA. ACM. 4.4.2

Yang, S.-H., Kolcz, A., Schlaikjer, A., and Gupta, P. (2014). Large-scale high-precision topic modeling on twitter. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1907–1916. ACM. 3.2.2

Yang, Y. and Eisenstein, J. (2015). Putting Things in Context: Community-specific Embedding Projections for Sentiment Analysis. *arXiv preprint arXiv:1511.06052*. 4.1.3

Yao, L., Riedel, S., and McCallum, A. (2013). Universal schema for entity type prediction. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 79–84. ACM. 3.1, 3.2.2, 3.4.2, 5.1

Zhai, C. and Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 334–342. ACM. 2.4.1

Zhang, M., Zhang, Y., and Vo, D.-T. (2015). Neural networks for open domain targeted sentiment. In *Conference on Empirical Methods in Natural Language Processing*. 4.1.2

ZHANG, Z., GENTILE, A. L., and CIRAVEGNA, F. (2013). Recent advances in methods of lexical semantic relatedness - a survey. *Natural Language Engineering*, 19(4):411–479. 5.2.3

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in r. *The Journal of Machine Learning Research*, 13(1):1059–1062. 4.4.2

Zhao, W., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. (2011). Comparing Twitter and Traditional Media Using Topic Models. In Clough, P., Foley, C., Gurrin, C., Jones, G., Kraaij, W., Lee, H., and Mudoch, V., editors, *Advances in Information Retrieval*, volume 6611 of *Lecture Notes in Computer Science*, pages 338–349. Springer Berlin / Heidelberg. 3.2.2

Zhu, M., Zhang, Y., Chen, W., Zhang, M., and Zhu, J. (2013). Fast and accurate shift-reduce constituent parsing. In *ACL (1)*, pages 434–443. 2.3