Thesis Proposal

Latent Cognitive Social Spaces: theory and methods for extracting prejudice from text

Kenneth Joseph

January 2015

Computation, Organizations and Society Program
Institute for Software Research
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.

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

Abstract

Prejudices define the biased views we hold of other people. Our prejudices play a role, both implicitly and explicitly, in every social situation we encounter. They tell us 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. Prejudices play a role in less mundane social processes as well. In particular, genocide is often motivated by negative prejudices of particular social groups.

Despite the omnipotent impact of prejudice on our lives, existent theory and methods used to understand it are lacking. Theoretical models tend to focus heavily on the cognitive or social dimensions of prejudice, rather than on a joint socio-cognitive explanation. Methodologically, research on prejudice is based largely on small-scale laboratory experiments and survey data that is costly to obtain.

The first portion of this thesis develops a new mathematical theory of prejudice that incorporates both its cognitive and social dimensions. The theory provides a parsimonious explanation of prejudice at the individual, group and culture-wide levels. The second part of this thesis develops two new tools to extract prejudices from existing text corpora, an approach that can provide broader data than laboratory experiments at a much lower cost than surveys.

The final part of this thesis applies the theory and tools I develop to two case studies. In the first, I use a corpora of Twitter data relevant to the Eric Garner and Michael Brown tragedies. I focus on providing a better understanding the "paradox of race" in America today, where few Americans are explicitly racist yet racial inequalities are as strong as ever. In the second case study, I use both Twitter and newspaper data from the "Arab Spring" in an attempt to better understand the socio-cognitive web of prejudices existent in the Arab World.

Contents

List of Figures

List of Tables

Chapter 1: Introduction

...the personification of the devil as the symbol of all evil assumes the living shape of the Jew.
Adolf Hitler

From an evolutionary perspective, it makes some sense that humans prefer the company of others who are similar to them. However, there is little reason to believe that we have evolved to *hate* those that are different from us (Brewer, 1991). Instead, it seems more likely that such hatred, and indeed many of our preferences for similar others, must be *learned*- over time, humans are "taught", in an informal sense, whom to like, whom to dislike and to what extent (Cikara and Van Bavel, 2014; Heise, 2007). Optimistically, the fact that our emotion-laden perceptions of others, or *prejudices* (Hewstone et al., 2002), are largely learned and not inherent in our DNA should mean that it is possible to change negative prejudices. However, empirical evidence suggests that we still have little idea as to how to go about doing so (Paluck and Green, 2009). As those attempting such interventions are realizing, while humans are not genetically predisposed to dislike people who are different, this does not mean that the cultural, social and institutional structures within which prejudices are engrained are readily subject to change (Dixon et al., 2012; Heise, 2007; Paluck, 2012).

At some point in my academic career, I hope to explore ways to effectively reduce prejudice within the constraints of a complex social system. Here, however, I consider more tangible concepts which are building blocks for this ultimate goal. This thesis focuses on three questions:

1. **How can the prejudices humans learn be modeled accurately and parsimoniously?** 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]. Can we simply say that school-children "dislike" LGBTQ individuals, or is there a better way to conceptualize such biases?
2. **How do we learn prejudices?** That is, what are the processes and the important factors that determine the prejudices of a given individual at a given time?
3. **What tools can we develop to extract, or "learn", current prejudices within complex socio-cultural systems?** Current methods for extracting existent prejudices are largely survey-based (Cuddy et al., 2007; Heise, 2010b). Is there a way we can leverage other data, specifically text corpora available on the web, to provide alternative empirical understandings of prejudice?

This thesis aims to advance current theory and methodology in order to address these three questions. In a theoretical vein, I expand upon *Affect Control Theory (ACT)* (Heise, 2007; Robin-

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

son et al., 2006), a prominent theory of affect, emotion and social interaction that provides critical insights into prejudice (Rogers et al., 2013). My expansion of ACT takes the form of a new theory I call *Latent Cognitive Social Spaces (LCSS)*. LCSS extends ACT primarily by adding an explicit cognitive component to the theory, specifically an associative (or *connectionist*) model of cognition. The addition of an associative cognitive model to ACT allows me to address three shortcomings of ACT as it currently stands.

First, while ACT includes perceptions of "traits" and "attributes" of individuals (Heise and MacKinnon, 2010), it does not concern itself with how perceptions of *cultural forms* (i.e. skills/ habits/values etc.; Miles, 2014), effectively more cognitive elements of culture, may have implications for prejudice. With ACT, we are unable to consider explicitly how, for example, the fact that "professors like classical music" may be useful in the study of prejudice. Second, ACT presents what I term a *first-order social effects model* in that it considers how social interactions change our prejudices of individuals who hold particular identities. However, "second-order effects" of social interaction on our perceptions of other, related identities are not modeled by ACT. Thus, ACT does not consider how information we learn about a particular identity (e.g. "hockey player") might "filter" to other, related identities (e.g. "athlete"). Finally, because ACT does not have an explicit cognitive model, it does not express how individuals use context to inform them of the identities that can arise in a particular social situation. Instead, explanations of the semantic relationships between identities (e.g. between the identities "doctor" and "patient") are offloaded to *institutions*, which "organize the huge number of identities that [one] can encounter" (Heise, 2007, pg. 28). I argue that such a model does not faithfully account for the ability of the human mind to act as an associative engine which itself retains much of the information ACT explains via institutions.

In Chapter 3, I describe how LCSS addresses each of these shortcomings of ACT via a sociocognitive framework that unifies perceptions of culture and identity, provides a better definition of the role of the institution, and that allows us to consider both emotional and semantic relationships between identities. Although LCSS provides these important extensions to ACT, the additions made are slight compared to the overarching framework that ACT scholars have developed over the past four decades. The most important element of ACT for my work is its roots in linguistic categorizations of people. ACT is thus well suited to the extraction of prejudice from text data.

Unfortunately, very little have been done to explore the use of raw text data to study ACT. To this end, this thesis presents two new Natural Language Processing (NLP) tools that leverage LCSS, and by extension ACT. The chief methodological contribution of this thesis is a Bayesian Network that can extract the cognitive and institutional representations of prejudice purported by LCSS from a text corpora and its associated meta-data. More specifically, the model I propose is interested in the following task:

> **Methodological Task 1 (MT1):** *Given a set of text data from a collection of individuals and meta-data (e.g. location) about these individuals, learn the following: 1) individual prejudices and how they change over time, 2) stable subcultural and global prejudices and latent 3) contexts and 4) institutions and how they affect the production of text*

The statistical model I develop is applied to three datasets comprising two thematic case stud-

ies. The first case study is conducted on data relevant to the "Arab Spring" (Gelvin, 2015). For this case study, I have one dataset drawn from Twitter[2], as well as a collection of approximately 600K newspaper articles. The second case study is conducted on data relevant to the Eric Garner and Michael Brown tragedies, which for purposes of brevity I will refer to as the "Ferguson" case study. Here, I currently consider only Twitter data. The Twitter dataset I use is a collection of approximately 1B tweets extracted from the full timelines of 1.2M users that were actively engaged in the discussion of these two tragedies.

A critical assumption of MT1 is, of course, that I can readily determine the identities and cultural forms that individuals hold prejudices of in the text data. While, as I will discuss, the process of extracting these entities from text is heavily related to other NLP problems, it is unique in that the set concepts desired are broader than the set of "Named Entities" in Named Entity Recognition yet narrower than the set of all "things" (e.g. nouns) that could easily be extracted using Part-of-Speech tagging. Thus, I develop a new methodology that considers the following novel task:

> **Methodological Task 2 (MT2):** *Given a set of text data, label each word in the text as being representative of a (possibly multi-word) identity, behavior, setting, individual or cultural form, or none of the above*[3]

The primary goal of MT2 is to inform the input representations of MT1. However, the identification of these entities from text can be used to answer three additional questions. First, as Heise and MacKinnon (2010) notes, one question of principle interest is simply to explore which identities are used in text- that is, can we produce a "survey of identities"? Second, we may ask, what are the relationships a culture perceives between identities, cultural forms, settings and behaviors, and how does the network of these relationships cluster into institutions? Finally, how, if at all, do the identities and institutions in a large cultural group vary over time and location? Where appropriate, I will consider each of these questions in my two case studies.

In the following chapters of this thesis proposal, I will provide more detail on the literature I rely on, the theory and methods I will develop, the datasets I will use and more details on what I hope to achieve in my two case studies. In the last chapter, I provide a statement of the contributions and limitations of the work I propose and a timeline for the completion of this dissertation. This document also contains two appendices. Appendix A provides a more complete literature review. Appendix B provides more details about my previous work that relates to the proposed efforts here.

[2]the specifics of which will be determined, as detailed in Chapter 6

[3]Note that the other entities in MT2 will be detailed later in this document

Chapter 2: Summarized Literature Review

In the interest of keeping this proposal to a manageable length, I here provide only a very short overview of the literature relevant to this thesis. A slightly more complete literature review that provides details on the points made here can be found in Appendix A. An even more complete review, a sizable portion of which has already been written, will be incorporated into the dissertation.

2.1 Relevant Theoretical Concepts

As suggested by Heise and MacKinnon (2010), *Affect Control Theory (ACT)* (Heise, 2007; Robinson et al., 2006) consists of three main assumptions, or components. First, it assumes a particular measurement system for the affective meaning of identities, behaviors, settings and modifiers, which I will collectively refer to as *entities*. Each entity is assumed to be represented by a particular linguistic form that has a specific, fundamental affective meaning, or *sentiment* (Robinson et al., 2006). These fundamental meanings are shared widely across large cultural groups, and thus an implicit consensus exists across this culture in the affective meaning of any particular entity. The affective meaning of a sentiment is defined in a three dimensional space with axes entitled Evaluation, Potency and Activity (*EPA*), each spanning the range of -4.3 to +4.3 (Osgood, 1969). The evaluation dimension describes the goodness/badness of an entity. The potency dimension describes the powerfulness/weakness of an entity, and the activity dimension describes the level of activity/passivity of an entity. Collectively, the position of an entity in this semantic space defines a *prejudice* (Hewstone et al., 2002; Rogers et al., 2013) of that entity held widely by a particular culture.

The second main component of Affect Control Theory is an empirical framework for how social events change our perceptions of entities. A *social event* is a social interaction in which an *actor* enacts a behavior on an *object*, perhaps within a particular setting (Heise, 2007, pg. 36). ACT's empirical framework defines how "pre-event" EPA ratings of the actor, behavior, object and setting change after observing the event. For example, a teacher should be seen as "less good" after beating up a child. The final assumption of ACT is a cybernetic control system which specifies how individuals will behave in particular situations. ACT is a "control theory" in that it assumes humans seek to maintain the fundamental, culture-wide sentiments of identities, behaviors and settings in transient impressions that are generated when social events are observed or carried out (Robinson et al., 2006). Humans will thus select actions that minimize the *deflection*, or difference, between their pre and post event perceptions of the identities, behaviors and settings involved in the event. The mathematics of this model are reviewed in Section A.1.2.

Beyond these three components, it is important to note that a distinction is made in ACT

between a "setting" and a "situation". A setting is a linguistic category that defines a place and time with a specific fundamental meaning. Situations, on the other hand, encompass the entire environment in which a social event plays out. Part of this environment is the set of contextual stimuli that individuals use to define their own identities and the identities of others around them. Context also provides cues for which *institution* is appropriate in a particular situation. Heise (2007) defines an *institution* as a "constellation of identities, settings, and actions relating to some elementary concern" (pg. 28). Institutions "organize the huge number of identities that you can encounter (pg. 28) and constrain the possible identities we can take on and the behaviors we are willing to enact.

I perceive three ways in which ACT as it currently stands could be extended to be even more useful in the study of prejudice[1]:

1. ACT is a *first-order social effects model* in that a social event impacts only the individual's perception of the identities and behaviors currently defined in the situation. It is also useful to consider *second order social effects*, where information learned from social interactions also affects our perceptions of closely related entities
2. ACT's definition of the institution and interpretation of how individuals determine appropriate institutions from contextual cues are intentionally ambiguous. In being so, ACT does not provide a formal link between context and institutional boundaries and ascribes too much power to institutions in defining appropriate identities for a particular situation.
3. ACT does not include perceptions of *cultural forms* into the model, where cultural forms are defined as more cognitive elements of culture such as skills, values and preferences Lizardo; Lizardo; Lizardo's (2006; 2011; 2014).

To address these limitations, I built an explicit model of cognition into ACT. Doing so requires the use of concepts from both cognitive psychology and cultural sociology. From cognitive psychology, I adopt schema theory and activation theory. Schema theory (Rumelhart, 1978) provides one model of cognition that explicitly details the associative nature of the human mind. Schema theory posits that associative memory can be modeled via *schemas*, which are labeled "data structures" that hold sets of variables which define what we expect from a particular entity Rumelhart (1978). Importantly, schemas are situated within networks, where links connect schema that our brains have come to believe are related to each other. Activation theory Anderson et al. (1997); Collins and Loftus (1975) details the process of how contextual stimuli *activate* schemas. Activation spreads from originating schema to other related schema through the process of spreading activation (Collins and Loftus, 1975). The same contextual stimuli may in time become associated with collections of schema through repeated simultaneous activation, thus leading to "institutional perceptions" within the mind.

From cultural sociology, I utilize elements of two different theories. First, I use concepts of Miles (2014) work on culture and identity. Because LCSS only loosely draws on this work, I leave a discussion of it to Appendix **??**. The second theory I draw on is Lizardo and Strand's (2010) Strong Practice Theory (SPT). SPT draws heavily on the work of Bourdieu (1986). It agrees with toolkit theory (Swidler, 1986) in that it considers culture to be more a set of heuristics that we use to inform our actions than some kind of knowledge "system" that we acquire through

[1]Note that in Appendix A, I address five issues, two of which have already been addressed by other researchers)

the socialization process. However, SPT departs from toolkit theory in that it believes individuals are capable of internalizing a rich and complex system of these practices and habits, rather than only internalizing a crude and shallow set of rules that we consistently apply. Lizardo and Strand (2010) argue that this internalization exists at the level of associative memory; in other words, that culture is systematic at the level of cognitive schemas. Associative knowledge encoded in institutions is thus often replicated within the mind. SPT thus gives a model for how to incorporate both ACT's model of the institution and the associative nature of the human mind into a single theoretical framework.

2.2 Relevant Methodological Concepts

In MT2, I will extract the entities mentioned above relevant to LCSS and ACT from text. The extraction of words, or sets of words, from a body of text that may refer to some concept of interest is known broadly as *entity mention detection*. Common tools utilized in models for mention detection include *part of speech tagging* (POS tagging) (e.g. Ratnaparkhi and others, 1996), a technique that determines the part of speech of every term in a sentence, and thesauri, or dictionaries, that map from various lexical *surface forms* into a single higher-order entity (e.g. Wikilinks[2]; CrossWiki[3]). Having extracted surface forms likely to refer to entities of interest, one may then be concerned with determining which real-world entity a surface form refers to (e.g. does "Michael Jordan" refer to the basketball player or the statistician?) and similarly, if two surface forms within the same text refer to the same higher-order concept (e.g. do "bike" and "bicycle" refer to the same thing?). These processes are referred to as *entity disambiguation* (Habib and Keulen, 2013; Moro et al., 2014; Wang et al., 2012) and *coreference resolution* (Lee et al., 2013; Soon et al., 2001), respectively. Additionally, one may be interested in attributes or categorizations of entities. The task of *named entity recognition* (NER) (e.g. for Twitter; Ritter et al., 2011) represents the extraction of concepts from text and their categorization into a general typology of named entities, most often people, places and organizations. The connection of entities to attributes of these entities from, e.g., Wikipedia (or more generally a *knowledge base* (Medelyan et al., 2013)) is known as the process of *entity linking* (Moro et al., 2014).

In MT1, I will be interested in semantic relationships that exist between entities. The process of *relation extraction* uses statistical regularities across many documents to extract typed relationships between concepts (e.g. `Obama` $\xrightarrow{\texttt{born\_in}}$ `Hawaii`) (Fader et al., 2011; Mitchell et al., 2015a). Most often, these approaches utilize some form of *dependency parsing* Kbler et al. (2009), which uses statistical models to determine linguistic dependencies between concepts within a particular text. Unfortunately, as David Bamman is considering in his thesis work (Bamman, 2014), the use of these types of relationships, often encoded in Knowledge Bases, may be problematic when studying culture and prejudice, as they may encode beliefs as undesirable, or even malicious, "facts" within the system (e.g. `Dogs` $\xrightarrow{\texttt{are}}$ `bad`).

Perhaps the best known method for determining relationships between concepts, however, is to simply use the co-occurrence of two concepts in the same textual unit (e.g. a sentence, paragraph or tweet) as evidence of a relationship. Direct utilization of these co-occurrence relationship is the basis for semantic network analysis Carley (1990); Carley and Kaufer (1993),

[2] `http://www.iesl.cs.umass.edu/data/wiki-links`
[3] `http://www-nlp.stanford.edu/pubs/crosswikis-data.tar.bz2/`

which is useful in understanding the "mental map" that authors of texts have between different concepts. Co-occurrence relationships in text can also be used to extract higher-order topics from text, where terms that repeatedly occur in the same text unit across many documents are assumed to be drawn from the same topic.

Perhaps the best known approach to extracting topics from text are unsupervised admixture models, known generally as *topic models*. The canonical example of a topic model is latent Dirichlet allocation (LDA; Blei et al., 2003). However, as discussed in Appendix A, much work has been done to extend LDA. Topic models are particularly useful in that they place text into a latent space with a constrained dimensionality. From this latent space, interpretation, in particular visualization (Dou et al., 2013), becomes much more straightforward. Such a space is precisely what the EPA profiles of ACT represent, and I will attempt to learn a similar embedding of concepts into a latent space with the model presented in Chapter 5.

EPA ratings of concepts are not only a latent space but also a model of sentiment. The extraction of the sentiment from 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 (Pang and Lee, 2008). The work I propose for my dissertation differs from most work on sentiment analysis in three ways. First, as opposed to estimating the sentiment associated with an entire text, I am much more interested in the sentiment associated with particular concepts (Cambria et al., 2014; Dong et al., 2014). Second, I am interested in individual and group level sentiments, rather than global perceptions by all individuals represented in the text (Hoang et al., 2014; Krishnan and Eisenstein, 2014; Mukherjee, 2014). Finally, while many sentiment analysis approaches evaluate concepts on a single, evaluative dimension, my work places concepts into a richer, more descriptive three-dimensional latent space. These efforts are in a similar vein to recent work by Kim et al. (2012), who learn a multi-dimensional representation of concept-level sentiment scores.

Chapter 3: Latent Cognitive Social Spaces

In this chapter, I describe LCSS as it currently exists, and then provide an outline of work still to be completed.

3.1 Overview of theoretical model

The theory of Latent Cognitive Social Spaces is defined using five sub-theorems. I state each in turn here, and then provide a summary of the important points implied by the model.

3.1.1 Subtheorems in LCSS

Individuals' schematic cognition can be represented by a set of attributed points in a Euclidean space

A particular individual's perception is represented as a set of schemas arranged within a schema network. In the LCSS model, these schema networks are represented as a Euclidean space. The use of Euclidean space as opposed to a network is chiefly a matter of convenience, as latent space network models show that the network and Euclidean space as relational representations are reasonably interchangeable (e.g. Hoff et al., 2002). I use Euclidean space as opposed to a network representation because of the ease of defining distances and positions in a Euclidean space, which will both be important in the mathematics of the theory.

In the LCSS model, individuals can have schemas of five types of entities: individuals, behaviors, settings, identities and cultural forms. Each schema is uniquely defined by a linguistic representation of the entity and has parameters that define the schema's*content*, its *position* in the Euclidean space and a *level of activation*. The content of a schema is defined to be a distribution over EPA space[1]. Schemas also have a distribution to represent their position in the Euclidean network. The activation of a schema is a single value.

Individuals use their LCSS to interpret social situations via a set of introspective *queries*. With respect to ACT, relevant queries may include, for example, "what identity should I take on?", "what is the identity of this individual?", or "what setting am I in?" While I will define the mathematics of several important queries during the thesis, I side-step these issues at present as they are not necessary to define the mathematics of the NLP methods defined in the later chapters.

The likelihood of an individual using a particular schema to help answer a particular query is

[1]Note that the idea of schema holding distributions as content is not entirely correct, at least not in the ACT-R model from which LCSS derives Anderson et al. (1997). However, the idea that schema are high-level cognitive entities that *blend* contents of individual memory chunks into distribution is roughly equivalent to the underlying model. I am indebted to Michael Martin for clarifying this

defined by the schema's level of activation. The activation level of a schema at any given time is based on three factors:

- Stimuli that "match" a schema cause an increase in the activation level of that schema. *Stimuli* are distributions over possibly unnamed, possibly contentless entities. That is, a stimuli may activate one or more entities based on name, or may activate a generic entity of a particular sort (e.g. an identity) with a particular EPA profile, or may provide both schema name and content[2]. Stimuli may be situation-specific (e.g. "hospital setting") or query specific ("what do I think of this individual, given that he is wearing a hat"?)
- Each schema has a *basal*, or resting, activation level Anderson (2007). Basal activation represents the activation "energy" that builds up over time- because activation decays exponentially, a schema that has been activated stays "active" for some time. Schemas that are frequently activated will have high levels of resting activation.
- Per spreading activation theory, the activation of a schema also prompts the activation of related schema, where the level of activation shared between two schema is inversely proportional to their distance in the Euclidean space.

Schema positions in the Euclidean space are thus important in determining the level of activation that spreads between any two schema. Because of this, and because the Euclidean space emulates network structure, the position of a schema is only interesting when it is given in relation to other schema. The more frequently two schema are instantiated in the same social situation, the stronger their relationship and thus the closer they will be in the latent space. Importantly, positions in the Euclidean space are thus distinct from positions in EPA space in that an individual can, for example, instantiate two competing role identities (patient and doctor) frequently together that have very different EPA profiles. The *semantic*, or *cognitive*, relationships provided by the Euclidean space form the basis for the cognitive replication of institutional structures.

Similar to schema content, schema positions are defined by a distribution. Variances in schema positions mediate distances between entities - in general, the more uncertainty there is over the position of an entity's schema in the LCSS, the weaker its association to all other entities. Uncertainty in schema position thus reflects the level to which an individual is confident in the associations between an entity and all others. This confidence should in general become stronger as an entity is activated frequently; thus variance in schema position should be inversely correlated with activation. However, this is not always the case - where schema are utilized in concert with a large number of other schema that are themselves unrelated , the schema may have a high variance in position but also a relatively stable level of activation. Conversely, schema may be infrequently instantiated, but always utilized in concert with one or two other very specific schema.

Each individual has a distribution of *attributes* that act as stimuli for perceptions of the individual held by others

Individuals are defined not only by their cognitions but also by a distribution over attributes, which are stimuli restricted to the space of identities and cultural forms emitted by the individ-

[2]Like Miles (2014), I skirt the means by which individuals come to associate real stimuli, e.g. those from the visual cortex, to entities defined in the LCSS. While Semantic Pointer Theory (Schröder et al., 2014) appears to be a useful tool to make this link in the future, I do not consider it here

ual. These attributes are similar to the physical features and subjective attributes of individuals defined by (Heise and MacKinnon, 2010, pg. 126). At any given time, an individual emits a single set of attributes, obtained via drawing a specific vector from their attribute distribution. This particular attribute set then acts as a stimulus for others inferring perceptions of the individual.

Some of the attributes in an individual's attribute set, like facial features, are difficult to change. Other attributes, like the clothing one wears, are much more malleable. In such cases, individuals have a choice as to which attributes they will present. The malleability of different attributes and how an individual makes decisions on which attributes to present are defined within a probabilistic structure in which schemas may inform the individual's attribute choices, and where attributes themselves may be correlated.

Contexts provide additional stimuli that activate schemas

At any given time, individuals exist within zero or more relevant contexts. When an individual is within a context, the context's stimuli are "applied" to the schema network of that individual. Contexts can vary in the extent to which they are available. Certain contexts are omnipresent - for example, national culture, while others, like sporting events, are temporary. Mathematically speaking, contexts can thus be defined by a distribution over the individuals, spatial regions and time periods to which they apply and by a particular set of stimuli.

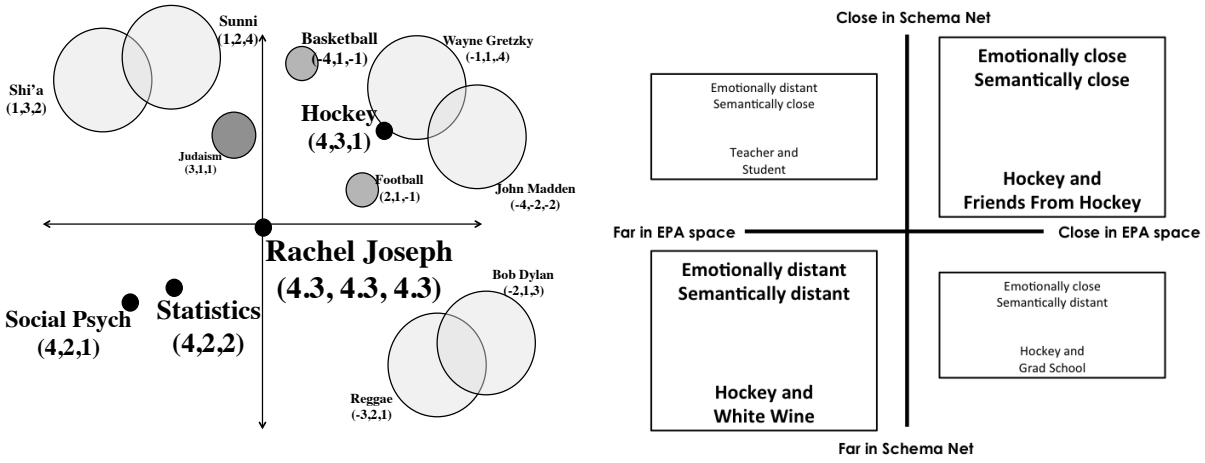
Institutions provide information and define and constrain perceptions

At any given time, agents may also have varying levels of access to institutions. Per SPT, institutionalized scripts for action may be used during social interaction to represent one's point of view. Per ACT, institutions may provide cues as to the "appropriate" identities, behaviors or cultural forms in the current setting. Both scripts for action and optimal/despised identities can be, and are, defined by an LCSS that represents the institution.

With respect to SPT, it is important to note that much of the information encapsulated in these institutionalized LCSS are also represented by individuals' cognitions. These links may be faint, requiring individuals to rely on institutions, or may be readily apparent to the individual. An important question thus becomes, when do individuals rely on institutional knowledge as opposed to their own, and when do particular institutions apply as opposed to others? These two questions are of critical importance to the incorporation of SPT into my theory, and are also important for how the cybernetic assumptions of ACT are modeled. Currently, my view is that the answers to these questions should be probabilistic. The closer an institution's LCSS matches the individual's LCSS, the more likely it is that it will be instantiated in that situation. Similarly, individuals may vary in their reliance on institutions in different situations, in particular when uncertainty is high. Combined, these likelihoods define the odds that an actor will utilize institutional knowledge in a particular situation. I do, however, plan to revisit these assumptions throughout my dissertation to provide more specifics as to this process, in particular by considering recent work from other scholars on the interrelationships between institutions, culture and prejudice (e.g. Ridgeway and Kricheli-Katz, 2013).

Individuals LCSS can be *transformed* by new information

When an individual is exposed to new information, it will transform their perception of the world around them. More specifically, information provided via interaction will cause schema positions, contents and activations to change. The activation of a schema (or creation, if the individual does not have a schema of a particular entity) and the updating of schema connections are defined via the mathematics of schema theory, activation theory and latent space network



(a) (b)

Figure 3.1: Sketches of the PGMs for static perception and learning

models. The updating of schema content is determined by the mathematical model imposed by Affect Control Theory, and more specifically via a synthesis of the Bayesian adaptations to the theory described by myself (Joseph et al., sub) and elsewhere in a similar fashion by (Hoey et al., 2013a).

3.1.2 Summary of Important Points

Figure 3.1a) provides an overview of the cognitive structure provided in this thesis to explain prejudices. Of course, such a structure is not (and is not intended to be) truly representative of cognition. Rather, it is a parsimonious model that I believe well represents many of the important features of the human mind at a level that suffices to explain prejudice. In the figure, each circle is a schema. Each schema is represented by a linguistic category placed either inside the circle or right next to it. Under each linguistic category, three numbers representing the content of the schema, its EPA rating, is given. The size and color intensity of each circle represent the certainty of the individual in the position of the schema in the LCSS. The size of the linguistic category represents the level of activation for each schema. The position of the schema in the Euclidean space, when compared to the positions of all other schemas, represents its semantic relationships. It should be noted, however that the dimensions of the Euclidean space, like in all latent space network models, are meaningless. The choice of utilizing two dimensions is explicitly made to encourage easier visualization of results. During my thesis, I will explore spaces with larger dimensions and weight the tradeoff between understandability and faithfulness to data when choosing the final dimensionality to be used for a particular dataset.

From the representation of a particular individual's cognition provided by Figure 3.1a), several pieces of information can be inferred. For example:

- The individual values the (other) individual "Rachel Joseph", as he frequently activates a schema for her and believes she is maximally good, powerful and active
- The individual perceives a strong but uncertain relationship between several concepts, including Bob Dylan and Reggae

- The individual has several clusters of entities he feels are related- in particular, it appears there are clusters for sports, music, school and religion
- Certain entities, such as basketball and hockey, are perceived by the individual as being semantically related but emotionally quite distinct

This toy example hints at the expressive power of LCSS. Perhaps most importantly, the individual can perceive two entities to be semantically similar or distinct and emotionally similar or distinct. This dichotomy creates four "quadrants" of entity similarity, portrayed in Figure 3.1b). In the top left, entities that are semantically close but emotionally distant should represent entities the individual often thinks about together, but in opposing emotional connotations. With respect to identities, this should represent opposing role identities ("teacher" and "student") and, perhaps, opposing social groups (e.g. "Sunni" and "Shi'a"). In the bottom right, emotionally close but semantically distant entities should be those that serve as important facets of the individual's perception in distinct contexts, a sort of representation of her "weak identities" (Smith-Lovin, 2007). On the diagonal, we see terms that are either close or distant in both emotional and semantic space at the same time. Although the emotional content of a schema is assumed to be independent of the emotional content of all other schema, Figure 3.1b) shows my belief that most terms should fall along this continuum. This reflects my belief that, due to the spreading of activation and the probabilistic nature of schema instantiation, similar schema should be activated in situations with similar emotional content. Thus, we should in general expect "good things" to be associated with other "good things", and the same for "bad things". Further, we should expect that perceptions of more specific identities, e.g. "hockey player", should by these same processes gradually *filter* to related but more generic identities, e.g. "athlete".

The above discretization of semantic and emotional closeness is important for two reasons. First, it provides an interesting avenue for hypothesis specification in the context of prejudice. In consideration of brevity, I here simply point out that this is particularly the case for the top left quadrant of Figure 3.1b), which should represent cases where an individual has polarized emotional views of two things that appear frequently in the same context. Second, Figure 3.1b) suggests the notion that when engaging in social situations, individuals must engage in two sorts of control processes, one semantic and one emotional. Thus, the deflection principle in LCSS can be stated as follows: *in the LCSS model, individuals seek to maintain both their semantic and emotional representations of the situation at hand.* This is the focal point around which the LCSS mathematical model is constructed.

In addition to this overview of the cognitive model of individuals posited by LCSS, it is also useful to quickly touch on the generative process by which a prejudice is assumed to be constructed. Figure 3.2 provides a sketch of a probabilistic graphical model that encodes the theoretical dependency structure induced by LCSS on the formation of a prejudice at time t. As the figure shows, the context at time t provides stimuli, which activate particular schema in the cognition of both the self (the individual who will hold the prejudice) and the "other" (the individual who will have a prejudice held against her). Both the self and the other's "previous cognition" at time $t-1$ influence the extent of this activation by defining the semantic relationships between schema. This prior cognition also largely defines the individual's emotional perceptions at time t. Given a perception at time t, the other draws a set of attributes, which provide further stimuli to inform the self. These stimuli, combined with contextual stimuli, the self's new LCSS at time t and information on desired and despised information from various institutions combine to

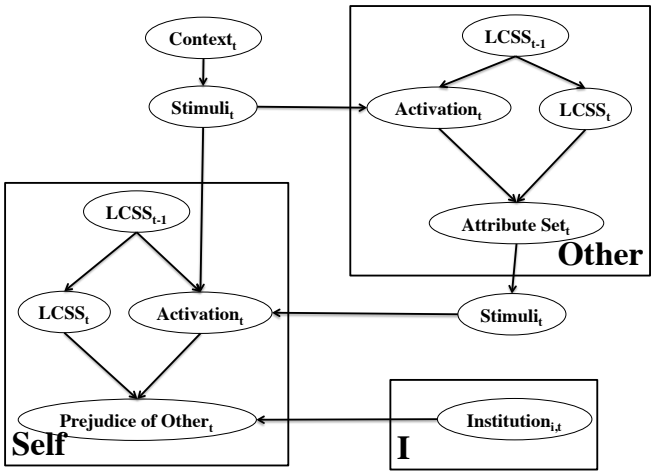


Figure 3.2: Sketch of the PGM for inference of prejudice

inform the individual of their prejudice of the other at time t.

3.2 Statement of Work

Notably absent from this chapter are the following, which will be completed for my dissertation:

- A description of the mathematical model. While previous work has led me to be well-acquainted with the mathematics of ACT (Joseph et al., sub) and schema and activation theory (Joseph et al., 2014c) independently, I have not yet completed the step of formally combining them into LCSS.
- At least the queries relevant to prejudice "how does person X feel about people of type Y?" must be mathematically formulated
- I will need to provide an example of how the model makes predictions akin to those made by ACT. In completing this portion, I will likely develop a toy agent-based model (Gilbert and Troitzsch, 2005) which I will use to show the importance of my work in comparison to a toy agent-based model model of ACT
- A concrete discussion of where predictions differ between LCSS and ACT, and where LCSS makes novel theoretical predictions of existing questions in the study of prejudice. While I touch on some of these points in my two case studies, a concrete discussion of the ways in which predictions differ between ACT and LCSS and more concrete ways in which LCSS extends the larger theory surrounding prejudice is still necessary. Some avenues that might be explored include:
  - The utility of LCSS in better predicting the object of a particular actor's behavior (by incorporating semantic knowledge) relative to ACT
  - The correlation between the semantic and emotional perception of relationships between entities and the meaning of outliers in this relationship
  - The relationship between uncertainty in semantic position and EPA ratings
  - The relationship between similarity in perception and connections in the underlying social network

Chapter 4: Extracting Identities, Cultural Forms, Settings, Behaviors and Individuals From Text

In Chapter 2 of their 2010 book, Heise and MacKinnon (2010) extract all terms from WordNet (Miller, 1995) that are lexical descendants (recursively) of the term "human being". The authors use the resulting 5,501 terms as an "identity survey" of the English language, using the hierarchy provided by the data to consider various properties about the structure of identity in the English-language "theory of people". In Chapter 3, the authors then take text from both an offline, professionally written dictionary and from WordNet term definitions and create a semantic network, where a link between two words exists if one word appears in the dictionary definition of the other. A clustering algorithm is then applied to the network, and the resulting clusters are described as the institutions governing the English language-speaking theory of people. These institutions incorporate identities but also settings and behaviors. For example, one cluster includes identities such as siblings and parents and activities such as caregiving, representing the institution of family and marriage (pg. 79)[1].

The approach taken by Heise and MacKinnon (2010) has certain advantages. The data used is relatively clean - both WordNet and the professional dictionary are human curated and widely used, suggesting a high level of precision. The cleanliness of the data also allows for a relatively straightforward analysis, one that can escape the complexities inherent in extracting such information from raw text at larger scales. However, the approach also has certain disadvantages. While the dataset used was curated by a very specific collection of (albeit very knowledgeable) individuals, the relations and definitions made by these individuals may suffer from a low level of recall in two ways. First, these individuals may not be aware of or consider as common knowledge additional aspects of the identity and institutional structures used informally by many English speakers. Second, their particular view may not be entirely reminiscent of sub-cultural views of identities, institutions or their relationships.

In this Chapter, I briefly describe how I will use Twitter, news media data and possibly Wikipedia data to construct several similar identity and institution surveys. The approach I develop will attest to the first limitation of Heise and MacKinnon's (2010) work above in that I will attempt to capture a broader array of LCSS entities than can be found in WordNet or in a conventional dictionary. My efforts will also extend the work of Heise and MacKinnon (2010)

[1]Note that the process by which Heise and MacKinnon (2010) come to these institutions is not unlike the process by which meaning is extracted from topic models.

by extracting cultural forms and learning their associations. However, the surveys produced in this thesis will have biases themselves, and thus will serve to compliment rather than to in any way undercut the analysis presented by Heise and MacKinnon (2010).

As I will discuss below, the efforts I present will also act as a compliment to a variety of recent work in the NLP domain. More specifically, I see my efforts as an application of a host of existing methodologies to a novel prediction problem.

4.1 Problem Description

4.1.1 Task

The focus of this chapter will be on constructing a model to accomplish the following task, repeated from Chapter 1:

> **Methodological Task 2 (MT2):** *Given a set of text data, label each word in the text as being representative of a (possibly multi-word) identity, behavior, setting, individual or cultural form, or none of the above*

Formally, I will develop a BIO tagger that identifies each term in the text as existing at either the *B*eginning of an LCSS entity, with*I*n an entity, or *O*utside of an entity. While the definitions of each entity are described in Appendix A, I review them below for clarity:

- An **identity** is defined as either a role, category or social identity. *Role* identities indicate positions in a social structure (e.g. doctor). *Category* memberships come from "identification with some characteristic, trait or attribute" (e.g. African American). *Social* identities indicate membership in social groups (e.g. Pittsburgh Steelers fan).
- A **behavior** is an action that one individual or identity enacts on another individual, identity or cultural form
- A **setting** is a linguistic category that define a place and time with a specific fundamental meaning
- An **individual** is a specific person entity, as defined in the task of Named Entity Recognition
- A **cultural form** is a skill, value or preference - a noun describing something that one can be good at, like, or prefer in comparison to another cultural form.

4.1.2 Input data and Feature Representation

The input data will be raw text, either from a tweet or a paragraph from a news article. However, features from prior tweets from the same individual or other text in the same article may be included.

Feature extraction will encompass Part-of-Speech tagging (Owoputi et al., 2013; Schmid, 1994), dependency parsing Kong et al. (2014); Kbler et al. (2009) and coreference resolution Lee et al. (2013); Soon et al. (2001), where output from prior and future words in the text will be used as features to predict the tag of the current word (in addition, of course to features of the current word). Various surface-form dictionaries, which map surface forms (e.g. "MJ") into real-world entities ("Michael Jordan") will also be used to define features. As possible, I will use information (i.e. entity class) of the real-world entities as further features where this information is available. For example, the CASOS Universal Thesaurus maps surface forms to nodes, which are given node classes (e.g. "agent" or "organization"). The WikiLinks dataset Singh et al.

(2012) similarly maps surface forms to Wikipedia pages, from which both structured and semi-structured data may be collected. Finally, the existing ACT dictionaries of identities, settings and behaviors (Heise, 2001) will provide useful information as well.

4.1.3 Model, Inference and Learning

I expect this portion of the methodological work to consist heavily of feature engineering and do not plan to develop a novel model. I will thus use off-the-shelf tools for inference and learning. The primary question is then, of course, what form of predictive model I will use. I plan to consider a variety of different approaches here and select the one which performs the best. In order to assess performance and to acquire data to train the model, I, along with other collaborators, will develop a gold-standard set of Twitter and newspaper data, on the order of 2,000 tweets and 2,000 news sentences.

I am still debating from what source these tweets and sentences should be drawn, in particular if they should be extracted from my case studies or from more generalizable pools of data. I am also still considering whether or not to ask annotators to differentiate between types of identities (i.e. role/social/category) and types of cultural forms (skills/values/preferences). If so, I would likely formulate a second tier prediction problem (i.e. given that X is an identity, what type of identity is it?) to predict these labels rather than incorporating this into the task as defined. Regardless, the gold standard data will be hand-labeled by at least two annotators, who will identify the (non-)existence of the LCSS entities of interest in each sample.

Having obtained a data source of gold-standard data, I first plan to treat the problem as a supervised learning problem, thus using a structured learning model to draw inferences. I may use a Conditional Random Field (Lafferty et al., 2001), as I am already familiar with them, but also will consider structured SVMs (Taskar et al., 2003) or more recent structured prediction models (Zhu and Xing, 2009). I then plan to extend this base model by considering a semi-supervised (Chapelle et al., 2006) approach in which I utilizing additional, unlabeled data to help inform the predictive model. Finally, the problem at hand seems ripe for distance supervision (Mintz et al., 2009), as we already have a significant amount of information about the terms of interest that may be readily used to provide noisy labels for non-gold standard data. While some of these labels may be incorrect (e.g. "police" is used as a verb instead of a noun, and is thus not an identity in the particular sentence), information leveraged from this noisy labeling process may still provide useful information to the predictive model.

4.1.4 Evaluation and Research Questions

The model will be evaluated on gold standard data. As I will be using off the shelf tools, I will not test the learning model itself. However, I will compare performance of my model to various other baseline approaches to evaluate its predictive ability. In particular, I will use the heuristic rules defined by Heise and MacKinnon (2010) to transfer from POS tags into the LCSS entity domain set and use this as a baseline model. I will also compare to a strict surface form mapping approach using, e.g., the ACT dictionaries, and also to a baseline approach that utilizes dependency parsing in addition to POS tagging, in combination with heuristic mappings into the LCSS entity set.

Provided the predictive model is reasonably accurate, I will use it to extract LCSS entities from the full set of data for each case study. In doing so, I will be able to ask the following

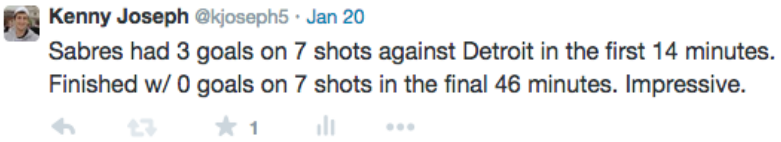


Figure 4.1: Example tweet

questions, in general, of at least one of my datasets:

- What are the identities extracted from the data and how to they compare to the extraction from Heise and MacKinnon (2010)?
- How does the full set of entities cluster, via co-occurrence relationships, into what Heise and MacKinnon (2010) would consider institutions?
- How do identity and institutional structures vary over time and space?

While I will make use of these questions to varying extents in the case studies in attempting to better understand the particular situations of interest, I will complete these steps for this section in order to get a high-level understanding of the usefulness of the methods developed. The estimation of this information will likely require additional statistical models to be developed. I do not believe, however, that these models will be complex enough to warrant discussion here, so I have omitted such a discussion from this document. It is also possible that I will make use of Wikipedia data as a comparative source for the identity and institutional structures uncovered in the data. Such work would be interesting, but possibly tangental to the overall goals of this thesis.

4.2 Expected Challenges

I see three chief difficulties in the implementation of this model:

- I will need a more concrete definition of cultural form before developing the gold standard data. For example, in looking at the example tweet in Figure 4.1, it isn't clear exactly what I would want to be extracted as cultural forms. In particular, the "Sabres" are a team in the National Hockey League. On the one hand, this information is useful, as it suggests I might enjoy hockey, and thus gives information about my preferences. On the other hand, whether or not the Sabres themselves are a "cultural form" is unclear. I will have to clarify such issues, perhaps by bringing in additional categories beyond the current set of LCSS entities. Otherwise, I suspect that both annotator agreement will be low and that there will be little information provided by the proposed features to help differentiate cultural forms from other nouns.
- From the perspective of analyzing results from the model, it is likely that many of the identities extracted as surface forms will refer to the same concept. It thus may be possible that entity disambiguation will have to be performed before analyzing the data. While it is also possible that incorporating disambiguation into the predictive model could improve results (Hajishirzi et al., 2013), I do not plan to attempt this during my thesis work.
- Both identities (Heise and MacKinnon, 2010) and cultural forms DAndrade (2001) have inherent hierarchical structure. While I defer consideration of the question of what the "right" level of hierarchy for the study of culture and social categories is, it may very well

be possible to infer hierarchical structure from co-occurrence and dependency relations in addition to the (flat) network structure that we can create from these two pieces of information as well.

4.3 Statement of Work

The following *must* be completed for this chapter:

- The creation of gold-standard data sets. I am still working out whether there will be one general Twitter and one general news gold standard data, or one per dataset used in my case studies.
- Feature extraction for the model
- A supervised prediction model that performs better than the mentioned baselines for *MT2*
- An exploration of semi-supervised advancements to this supervised prediction model
- The development of a method to extract an "identity survey" (Heise and MacKinnon, 2010, Ch.2) and the use of this method on at least one dataset
- The development of a method to extract an "institutional survey" (Heise and MacKinnon, 2010, Ch. 3) for at least one dataset

Other tasks proposed in this chapter, in particular distance supervision-based extensions to the predictive model and comparison to Wikipedia data, are not considered necessary tasks but may be completed as deemed necessary by further work.

Chapter 5: Extracting Latent Cognitive Social Spaces from text

The work performed in Chapter 4 will allow me to uncover entities important in LCSS. In this chapter, I develop a new methodology that allows me to infer individual, (sub)cultural and institutional prejudices of these entities. The primary assumption in my model is that each text is generated by an individual who is engaging in a social event (with a possibly imagined audience Marwick and Boyd (2011)) or who is describing a perception of a social event. LCSS posits that these interactions are informed by the individual's LCSS, and thus the text written presents us with information we can use to infer properties of the individual's cognition. The methodology I describe draws on my prior work extracting EPA profiles of identities and behaviors from the Arab Spring newspaper corpus. More detail on this work can be found in Appendix B.

5.1 Model Overview

5.1.1 Task

The model I present has the goal of inferring four different things:

- Individuals' sentiments towards and perceived relationships between entities and how these perceptions change (or do not change) over time
- Contexts in which texts are sent and how these contexts affect specific texts
- Institutional knowledge structures and how they affect different actors to varying extents
- How attributes of individuals may change particular prejudices; in other words what subcultural differences exist within the data given the attributes provided by the metadata

5.1.2 Data Representation

The chief concerns for data representation are how to represent an actor and a particular text. Each actor a is represented by a vector of binary attributes. The specific attributes an actor has will be determined by the dataset. For my case study on the Ferguson data, I plan to use discretized sociodemographic characteristics of the individual as well as a discretized version of the locations (e.g. states) that they tweet from as attributes. These will be inferred using techniques as described in a variety of recent work (Compton et al., 2014; Flatow et al., 2014; Mahmud et al., 2014; Mislove et al., 2011; Wagner et al., 2013). For the Arab Spring Twitter data, the meta-data of interest is simply the country or countries that the user is either located in or tends to focus on. For the Arab Spring newspaper data, I restrict exploration of the meta-data to the (discretized) location of the news agencies from whom the articles were extracted.

With respect to the text, LCSS is based largely on the idea that individuals have both per-

(a) (b)

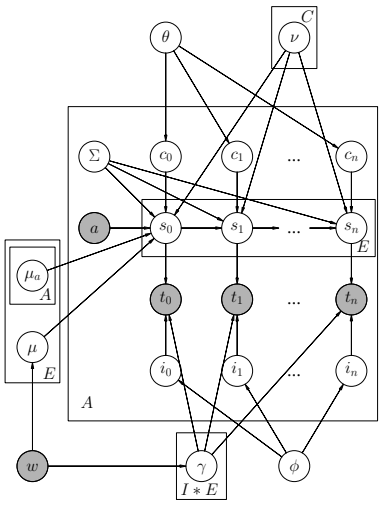
Figure 5.1: Two example tweets

ceptions of semantic relationships between entities and sentiments towards entities. When an individual is expressing an opinion or stating a fact, as in the tweet in Figure 5.1a), the question of how to represent a text becomes a question of how to extract and model relationships between entities and then, independently, how to extract and model the sentiment that the text displays towards these entities. Complications arise when individuals discuss social events, either those they have engaged in or events conducted by other individuals. An example of such a text is given in Figure 5.1b). In such a case, ACT informs us that, given information about the individual's perception of Chicago religious leaders and protests, we should be able to infer information about how the individual feels about Ferguson. Thus, both sentiment and relational information exist within this social event. Unfortunately, the sentiment information for any one element of the event is conditional on the individual's perception of the other two elements.

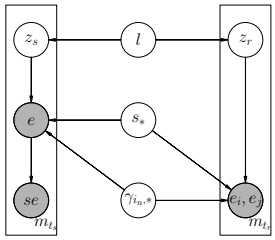
Each text, t, can thus be represented as three pieces of information:

- t_r is the set of dyadic $r$elationships between entities found in the text, extracted, e.g., via co-occurrence
- t_s is the set of entities for which a $s$entiment is expressed, along with their corresponding sentiment values. This is extracted via, e.g., assigning all entities in a text the sentiment score that a simple sentiment mining tool applies to the entire text
- t_v is the set of social events, where a social e$v$ent includes one entity enacting a behavior on another entity. These may be extracted via, e.g., heuristic rules applied to a dependency parse.

As discussed in Chapter A, there are a host of ways to extract relationships, emotional perceptions and social events from text. With respect to relational extraction, simple methods, such as considering all entities in a particular text to be related, can provide useful information. However, more complex approaches that consider lexical dependencies between terms (Kbler et al., 2009) or even external knowledge in creating connections between terms (Mitchell et al., 2015b) can provide significantly more nuance at the cost of increased computation and algorithmic complexity. Similarly, with respect to sentiment detection, it is rather straightforward to estimate the general sentiment of an entire text or tweet, but research on extracting sentiment towards a particular entity within a text is in its infancy. Social event extraction by default requires the use of at least dependency parsing, however, more complex approaches certainly exist. In this thesis, I will start with the simplest possible approach for relation, concept-level sentiment and social event extraction which gives me qualitatively reasonable output and which allows me to perform well on the evaluation tasks explored below. As necessary or feasible, I will refine these approaches to improve my understanding of the data and accuracy on evaluation tasks.



(a)



(b)

Figure 5.2: Graphical representation of the proposed model

Table 5.1: Variables used in the proposed model

Var.	Description
θ	A distribution over contexts. There are $\|C\|$ such contexts
ϕ	A distribution over institutions. There are $\|I\|$ such contexts
a	The attributes of a particular actor. There are $\|A\|$ actors in total
$c_{a,n}$	The context in which the nth text was generated for actor a
ν_c	A vector of length E that represents the extent to which a context c activates the schema for each entity
$\gamma_{i,e}$	The schema for entity e for institution i
$i_{a,n}$	The institution governing the nth tex for actor a
$s_{a,n,e}$	The schema s for actor a for entity e text at time n. The variable $s_{a,n,*}$ is the full perception of a at n
μ_e	The prior (fundamental meaning) for a schema of entity e
μ_{e,a_i}	The change in the prior (fundamental meaning) for a schema of entity e if an actor a has attribute i
Σ_a	The covariance matrix for all schema s at all times n for an actor a
$se_{a,n,e}$	The sentiment of the nth text generated by a for entity e There are $\|T_a\|$ texts per user
w	Institutional knowledge extracted from knowledge bases
A	Transition matrix from $p_{n,s}$ to $p_{n-1,s}$
f	A function that provides a Multinomial distribution over the likelihood of all pairs of schema
l_a	Likelihood for actor a that a particular sentiment or relationship is drawn from an institutional vs individual perception
q_r	Parameter governing the number of relationships emitted in any particular text
q_s	Parameter governing the number of sentiments emitted in any particular text
b	is a function to convert a particular perception distribution into a single vector of multinomial probabilities.
z_s, z_r	determine whether a particular sentiment or relationship (respectively) is drawn from institutional or individual perception

5.1.3 Model

A visual description of the statistical model I will develop is presented in Figure 5.2a) using a probabilistic graphical model (Koller and Friedman, 2009); the variables that are used are described in Table 5.1. A more detailed view of the representation of a text is provided in Figure 5.2b) and is discussed in more detail below. In all cases, I have omitted the various priors placed on model parameters; with space constraints in mind, I largely omit a discussion of priors here. For purposes of introducing the model, I will focus on its use with Twitter data, where actors are individual users and texts are tweets. In the case of newspaper data, the analog of a user is an individual news company (e.g. the New York Times), and the analog of a text is a single newspaper article.

Inference is performed for all entities in a set E, which are extracted from a set of texts, T, via the process described in Chapter 4. The texts in T are emitted by a set of actors, A, who are defined by a set of attributes, over a set of time periods, N. In addition, I assume a set of (latent) contexts C within which all texts are generated, and a set of partially observed institutions, I that also affect the generation of text. Each actor is also assumed to have a (latent) perception, s_a at

time n. I will here focus on a single actor, a, and thus drop all a subscripts to reduce notational complexity.

Each individual's perception is a set of points drawn from a multivariate Gaussian distribution with six dimensions. Each point represents a schema for a particular entity e. Thus, each perception is made up of $|E|$ points. The six dimensions of the Gaussian distribution are for the position of the schema in the two-dimensional latent space, the evaluative, potency and activity sentiment dimensions that make up the content of the schema, and the current activation level of the schema. Note that the variable $s_{n,*}$ refers to the full set of points/schema for the actor at time n. This set collectively defines the actor's *perception* at time n.

Actor's initial perceptions are influenced by two sources of prior information. First, the variable μ_e holds the fundamental (global) meaning, initial position and initial activation level for the schema $s_{0,e}$. In general μ_* holds priors for $s_{0,*}$. The variables μ_{e,a_i} holds the change in the fundamental meaning for the schema of entity e that can be expected if an actor holds the attribute a_i. The assumption that actor attributes affect fundamental meanings captures the belief that users with similar attributes are likely to share similar initial perceptions because they are likely to have been placed within more similar contextual and institutional structures and to have had interactions with more similar others prior to the data we observe (Stryker, 1980).

An actor's initial perception for a particular entity e is thus defined as $s_{0,e} \sim \mathrm{N}(\mu + a^T \mu_a, \Sigma)$, where Σ is the covariance matrix for all schema for an actor $s_{*,*}$. Note that, while not pictured in Figure 5.2a), hyperparameters on either μ or μ_a, or some sort of regularization term on μ_a in the objective function, can all be set in order to specify the extent to which attributes are expected to mediate fundamental sentiments. Further, a prior on Σ can be implemented to determine the extent to which individual meanings vary from meanings induced by the priors μ and μ_a.

Beyond this initial time point, each actor's perception over time is modeled as a Markov chain, where the location of a particular schema at time n, $s_{n,e}$ is conditionally independent of all other previous time periods given the schema's definition at time $n-1$, $s_{n-1,e}$. A transformation matrix A determines the relationship between $s_{n-1,e}$ and $s_{n,s}$. The matrix A is a possibly non-linear transformation matrix that may either be learned or assumed and tuned via cross-validation with respect to the dynamics model.

At time n, the actor's perception $s_{n,*}$ is also affected by the particular context c_n in which the actor is situated. This context indicator is drawn from an overarching distribution θ. A context (as opposed to a context indicator) ν_c is a vector of length E that represents the extent to which a context activates the schema for each entity. In addition to representing context, the proposed model also has a representation of institutions. More specifically, each tweet a user sends is influenced by a single institution i, drawn from a distribution ϕ over all institutions. Representations of institutional knowledge is held in γ. The institutional knowledge structure for a particular institution i is held in $\gamma_{i,*}$. Thus, the variable $\gamma_{i,e}$ refers to the institutional schema for entity e, and the variables $\gamma_{i,*}$ and $s_{n,*}$ are of equivalent forms. Note that institutional structures are assumed to be static. While institutional knowledge may change over time, I assume that the knowledge encoded within them is static over the duration of each dataset studied, none of which last longer than four years. Importantly, this does not mean that individual's use of particular institutions is static, and thus individuals may have, for example, stopped "using" government institutions during the Arab Spring.

The variable γ is influenced by w, which represents known institutionalized knowledge that

I will provide for the model from external knowledge sources. These knowledge sources may include Wikipedia and the institutions defined by INTERACT, a Java program that implements the ACT mathematical model (Heise, 2007). At a theoretical level, LCSS posits that this knowledge can be expected to influence how users express themselves in text. At a methodological level, this knowledge can significantly enhance the model's ability to predict what users will say. Thus, certain institutions will be represented by strong priors from w, while others may be more loosely tied to this knowledge and thus be able to extract latent institutionalized knowledge that the model infers.

Both an actor's perception at time n and the institutional knowledge governing the text at time n influence the text, t_n, that the actor emits. As described above, a text t is comprised of three sets. The generative process for a particular text is expressed in Figure 5.2b). As I am still working out how to handle social events, I will only consider t_s and t_r here. These two variables are sets, and thus have variable sizes. The distribution over sizes of t_s and t_r, defined by auxilliary variables m_{t_s} and m_{t_r}, are determined by two global parameters q_s and q_r[1]. With respect to the elements within these sets, the likelihood of any particular relationship or sentiment in a text t at time n is a function the individual's perception $s_{n,*}$ and the institutional perception $\gamma_{i_n,*}$. The variables z_s and z_r, drawn from a single, static Bernoulli distribution with parameter l, determine whether a particular sentiment expression or relationship, respectively, is drawn from an institutional or individual perception.

For any given relationship between two entities e_i and e_j, the likelihood it is emitted in text t_n is determined by the function $f()$, which incorporates both schema position and schema activation and provides a Multinomial distribution over the likelihood of all pairs of entities. The likelihood of a sentiment being expressed for a particular entity can be decomposed into the likelihood that any particular entity's schema is chosen, determined by the function b, and the likelihood of a particular sentiment value $se_{n,e}$ for that entity. The function b uses activation theory to convert a particular perception distribution into a single vector of multinomial probabilities. In the latter, note that $se_{n,e}$ is actually three dimensional - for notational ease, I have removed subscripts here, but the marginal of these three dimensions is still trivially a multivariate Normal distribution.

The model as presented leads to the following generative story for any given text sent at time n by user a:

1. Pick a context indicator $c_n \sim \mathrm{Cat}(\theta)$
2. Draw an institution indicator, $i_n \sim \mathrm{Cat}(\phi)$
3. For each entity e, draw a schema $s_{n,e} \sim \mathrm{N}(A s_{n-1,e}\, \nu_{c_n}, \Sigma)$
4. Draw a number of entities to emit sentiments for $m_{t_s} \sim \mathrm{Pois}(q_s)$.
   For each:
   - Draw institution or individual indicator $z_s \sim \mathrm{Ber}(l)$
   - Draw entity $e \sim \mathrm{Mult}(b(s_{n,*}^{z_s} + \gamma_{i_n,*}^{1-z_s}))$
   - Draw sentiment for e, $se_{n,e} \sim \mathrm{N}(s_{n,e}^{z_s} + \gamma_{i_n,e}^{1-z_s}, \Sigma)$
5. Draw a number of entity relationships to emit $m_{t_s} \sim \mathrm{Pois}(q_r)$.

[1]Similar to the original LDA article, where N is drawn from a Poisson but implicitly assumed in the graphical model, I do not display q_s or q_r in Figure 5.2b)

For each:

- Draw institution or individual indicator $z_r \sim \text{Ber}(l)$
- Draw relationship $e_i, e_j \sim f(s_{n,*}^{z_r} + \gamma_{i_n,*}^{1-z_r})$

5.1.4 Inference and Learning

I do not know exactly how inference and learning will proceed. However, my previous work and work implementing the model from the vaguely similar problem explored by Eisenstein et al. suggests that a piece-wise, Monte-Carlo Expectation Maximization (MCEM) framework is appropriate (Bishop and others, 2006). Using a piece-wise MCEM approach for inference and learning is particularly attractive for two reasons. First, if done correctly, I believe that sampling for each user in the Expectation stage may be done in parallel, greatly reducing computational costs. Second, this approach may allow me to develop an iterative solution to the problem of utilizing information from social events, an approach which seems natural given the problem.

With respect to the Monte Carlo procedure in the Expectation step, a variety of tools from Bayesian computation will need to be used. In particular, as the model specifies an autoregressive model with a non-Gaussian emission probability for s, I will need appropriate methods to sample from Markov models, in particular Forward-Filtering Backwards Sampling (Godsill et al., 2004). In addition, I will take cues from the literature on Bayesian estimation of latent space network models (Hoff et al., 2002) in order to develop the semantic relational model for both s and i.

5.1.5 Evaluation

Three types of evaluation will be performed:

1. I will ensure that the inference algorithm I develop is learning the "right thing". In other words, I will run my estimation model on simulated data to ensure that the model can recover parameters specified in the simulated model correctly.
2. I will consider how well my model is able to learn a representation of the data that is "better" than other baseline models. The baseline models I will compare to will be a model that ignores relationship information between entities (i.e. "vanilla" ACT), a model that ignores sentiment information (i.e. a connectionist cognitive model) as well as to at least one or a set of other models in the literature that can make either sentiment or relational inferences from text, or both. I will compare my approach to these baseline models in one or more of three possible ways:
   (a) Train the model on a subset of actors and calculate perplexity on held-out actors
   (b) Train the model on all actors and calculate perplexity on a set of held-out texts from these actors
   (c) Try to predict the sentiment of all actors towards a particular entity. Here, I would consider important entities, e.g. the police in the Ferguson data.
3. I will visualize model output and try to understand how results fit into contemporary understandings of the factors underlying the two case studies of interest. The model will be successful in this vein if it can provide visualizations that confirm expert opinions and provide new information that we can use to reinterpret or provide new explanations for the events in Ferguson and the MENA region. In general, it will also be successful if the visualization can provide a useful summarization of model output to those unfamiliar with

the statistical model but who are interested in LCSS as a theory.

5.2 Expected Challenges/Possible Additions

The primary challenges will, I expect, surround efficient and effective estimation of the model, which will likely include both tweaking the current specifications and being intelligent in the algorithms utilized. With respect to tweaking the model, it may make sense to treat perceptions, $s_{n,*}$ as mixtures of Gaussians rather than as drawn from a single Gaussian in order to "encourage" points to organize into coherent sub-groupings. It may also make sense to allow for a subset of actor attributes to be latent, in a sense creating groups based on similar perceptions. Beyond this, I will also have to address the following challenges:

- As "distances between a set of points in Euclidean space are invariant" (Hoff et al., 2002), I will have to take some measure to ensure positions of entities in schemas of different individuals are comparable
- How I choose to extract, represent, model and thus utilize information from social events found in the text
- Extraction of institutional knowledge from knowledge bases and how best to incorporate this into the model
- Per-concept sentiment extraction
- The model may be over-specified in that attempting to learn contexts, institutions and individual perceptions may lead to too many parameters to learn for the data I have or to too many configurations that achieve similar likelihood values to actually be able to interpret model output

5.3 Statement of work

The following tasks that have been outlined in this chapter must be completed for my thesis:

1. A method to extract relationships between and sentiments of entities within a particular tweet or newspaper article
2. The extraction of at least location information for users and/or newspaper articles in my datasets
3. The development of efficient algorithms for learning and inference of the model presented
4. An evaluation of this algorithm developed with simulated data
5. A comparison of the predictive ability of my model versus several baseline models
6. An evaluation of how well the output of the model can answer questions I will pose about my two case studies and the questions posed in Section 3.2 about LCSS
7. Implementation of a methodology to extract institutional knowledge structures from Wikipedia and the ACT dictionaries and incorporation of output into the model

Importantly, the model as described here is what I would consider a "full model"; it encapsulates a large number of the assumptions of LCSS. It is both impractical and illogical to begin with this full model, in particular due to the possibility that I will not even be able to estimate this full model with the data I have, as mentioned above. Instead, I will build slowly up to this full model, beginning first with the estimation of static perceptions of individuals only, and then adding in institutions, contexts and structured knowledge from institutions.

Chapter 6: Case Study 1 - "Arab Spring"

On December 17th, 2010, Mohamed Bouazizi immolated himself in Sidi Bouzid, Libya in response to harassment from both a local policewoman and local municipality officers. Bouazizi's case resounded with others who took to the streets in protest of constant harassment and victimization by a corrupt government. Although early protests were relatively small and were met with violence from government forces, social media sites like Twitter, Facebook and YouTube were used to record these events and display them to the broader public. These events are widely considered to be the beginning of what has come to be known, for better or worse (Gelvin, 2015), as the Arab Spring.

These events suggest that social media, and new media (Baym, 2010) more generally, played an important role in Arab Spring. However, it is almost common knowledge at this point that popular emphasis on social media as *the* cause of the revolutions is overblown (Bruns et al., 2013; Comunello and Anzera, 2012; Goldstone, 2013). Recent research has thus instead focused on how social media may have aided certain aspects of the revolutions in important ways for different people (Gall et al., 2013; Lotan et al., 2011; Starbird and Palen, 2012; Tufekci and Wilson, 2012), and in how social processes that were carried out via new media are reflective of those that occurred "offline" (Comunello and Anzera, 2012). Similarly, recent work, including my own, has suggested that data from newspaper articles written during the time of the Arab Spring also may be of use in better understanding these processes (Joseph et al., 2014a; Pfeffer and Carley, 2012). Thus, social media and the coverage of news media should be seen as both pieces and reflections of a complex system of causal structures that were at play. Elsewhere, I am exploring these causal factors using other NLP techniques, in particular by using the model presented by Eisenstein et al..

A host of historical factors played causal roles in the Arab Spring[1]. Perhaps the most important historical issues were long-standing economic woes, government oppression and social divisions within the Arab World (Comunello and Anzera, 2012; Goldstone, 2011). Governments in the Arab world are routinely ranked as some of the most oppressive governments in the world (Gelvin, 2015), restricting the civil and social rights of their citizens and clamping down on civil unrest swiftly and violently. Such actions are, in many countries, disproportionately geared towards specific ethnic or religious groups, as ethnic divisions in the Arab region are particularly salient, complex and violent. Members of different sects of Islam routinely carry out terrorist activities against other sects. Further, different nations in the region have different ethno-religious make ups, which means that different groups are oppressed in different countries and that national, in addition to ethnic tensions, are heated. Finally, while international perspectives on the

[1]For a more detailed review, see (Gelvin, 2015)

region vary, it can generally be stated that the international community at large holds a strongly negative prejudice of many of the ethnic, religious and national identities in the region.

The Arab World thus make a particularly interesting case study of prejudice in a dynamic socio-cultural environment. Further, the Arab Spring presents a particularly interesting time period to study, as group and role boundaries became even more salient and government and social institutions were overthrown and/or recapitulated. Unfortunately, much of the textual information relevant to native Arab's social perceptions and how they changed is, of course, in Arabic. Because I will not develop methods for Arabic text, I currently plan to focus largely on the international community's perception of the Arab world[2]

While it would be nice to study native perceptions in the region, and this may be possible to a certain extent, several interesting questions can still be asked of the international perceptions of the region. First, I can use the methodology described in Chapter 4 to created an identity survey that will help to explain which entities the international community tended to focus on and to group together into institutional clusters. Using this identity survey, I can also ask, how do these perspectives change based on the country of origin of the Twitter user (where available) or newspaper agency, and on what country the tweet or article is focusing on? Finally, using this data we can also consider how the identities focused on in newspaper data and Twitter data may differ.

I can also use the model described in Chapter 5 to extract fundamental sentiments held towards national and prominent ethnic identities in the region and consider effects of user/agency location and the media used. This is essentially an analysis of EPA ratings and how these differ across subcultures (as defined by user/agency attributes). In my prior work, I uncovered the existence of a discrepancy in EPA ratings of different Muslim identities in the news articles in the data used here. While the Muslim identity was viewed as relatively neutral, both the Sunni and Islamist identities, each representing social groups in which all members are practicing Muslims, were viewed much more negatively.

What my previous work does not consider, however, and what I will focus on in this case study, is how the international community viewed the semantic relationship between the Sunni, Islamist and Muslim identities, as well as between other entities of interest. LCSS posits that there should be a correlation between semantic and emotional proximity in perception. Outliers from this correlation should represent competing role identities or social groups. Thus, the question becomes, what is the semantic relationship between Sunni, Islamist and Muslim identities in the data, and how does this relate to their emotional relationship? More generally, do semantic and emotional relationships in general share the expected correlation? And do semantic and emotional relationships between competing social groups, in particular Shi'a and Sunni Muslims, actually fall into the "off-diagonal" of unreciprocated emotional and semantic proximity?

6.1 Data

Both the Twitter and the newspaper data I collect are focused on sixteen of the eighteen countries pictured in Figure 6.1. While we attempted to collect data for Western Sahara and Qatar, such data was not available in a large enough quantity to be of practical use. Note that while I generally

[2]While I do know that around 150,000 Twitter users sent one or more English texts from a location inside a country within the Arabic world, the newspaper data is certainly not extracted from within the region.

Figure 6.1: Arab Spring countries

will refer to these countries as consisting of the Arab world, a more accurate depiction is as a subset of nations in the MENA region, as Iran is not a member of the Arab World.

6.1.1 Newspaper data

The Arab Spring newspaper data was collected by first pulling newspaper articles from major English-based world news drawn from LexisNexis Academic. LexisNexis uses a proprietary algorithm to associate each article with a set of index terms. We query the LexisNexis database for any articles indexed by one of the eighteen country names shown in the legend of the map in Figure 6.1. Further details about the newspaper articles can be found in my previous work, (Joseph et al., 2014a). In total, I have approximately 600K newspaper articles from approximately 300 different newspaper agencies.

6.1.2 Twitter data

The Arab Spring Twitter data was collected from two sources from April 2009 to November 2013. 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 in Figure 6.1 above.

The above dataset consists of approximately 81M tweets. For my thesis, I have 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 constitutes around 11.2M tweets from around 300,000 users. Although this is a large quantity of data already, it may make sense to add additional data to the collection by using the same process described in Section 7.1 to "fill in" additional data about the users. If I do choose to do so, it should not affect my timeline in any critical way. I already have developed the technical infrastructure to collect the timelines and follower/followee relationships of approximately 30,000 users per day, and used these tools to collect data for the Ferguson case study.

6.2 Statement of Work

As the Arab world presents an intricate collection of cross-cutting (Blau, 1977, 1974; McPherson, 2004) national, regional and ethnic identities, I expect that additional questions beyond those posed above will arise during my thesis. The temporal dynamics of the Arab Spring, the fall of institutions and the possibility of uncertainty in the international community towards particular identities (e.g. the Islamist identity) all may provide avenues to explore the insight that LCSS does or does not allow for. Regardless, the pieces of work I will definitely complete during my dissertation are the following:

1. An "identity survey" on both datasets described here, using the methodology described in Chapter 4
2. An estimation of the statistical model in Chapter 5 on both datasets described here
3. A comparison of the semantic relationships between national and ethnic identities between news agencies and between individuals from different nations, and across the news and Twitter datasets in aggregate
4. A comparison of the prejudices held towards national and ethnic identities between the two datasets and within the two datasets based on user/news agency location
5. How prejudices and semantic relationships changed over time and how this may help to explain the success and/or failure of different revolutions in different nations during the Arab Spring

Chapter 7: Case Study 2 -“Ferguson”

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 chooses 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 Michael Brown and Eric Garner tragedies certainly were used as fodder for racial comments online. However, these events angered far more American citizens, both black and white, than they pleased. How could something like this happen in today’s America, many asked, where racism is all but extinct (Harris and Lieberman, 2015)? This question typifies the“paradox of race” (Williams, 1997) in America today. On the one hand, the vast majority of us do not hold, or at least do not claim to hold, any racial biases. On the other hand, differences between the quality of life of black and white Americans are as large as ever. This ‘racism without racists” (Bonilla-Silva, 2006, c.f. (Harris and Lieberman, 2015)) may be perpetuated in part by the following five interrelated factors:

- Many Americans still believe that the general public holds negative racial stereotypes, even if they themselves do not (Fiske et al., 2002; Hall et al., 2015)
- Institutions perpetuate negative prejudices and racial inequalities. For example, the prison system does so by disproportionately imprisoning African Americans Saperstein et al. (2013)
- While individuals may not consciously hold prejudices, our subconscious holds implicit biases we do not even realize we have(Avenanti et al., 2010; Cikara and Van Bavel, 2014; Van Bavel et al., 2008). Situations that impose threats or fear may encourage subconscious prejudices to become conscious ones (Cuddy et al., 2007; Smith-Lovin, 1987)
- Individuals may hold “indirect” prejudices, in which negative attitudes are expressed for entities other than race but that are disproportionately associated with a particular race.

Thus, Americans may indeed not be racist but rather be "racist by cognitive association". Alternatively, Americans may indeed hold racial prejudices but simply explicate these prejudices via more institutionally acceptable actions or concepts (Lizardo and Strand, 2010)

- Americans may simply be apathetic, only willing to address inequality in superficial ways (Christensen, 2011)

On one level, the paradox of race is explained rather easily by Affect Control Theory. Under the control principle, and assuming Americans do indeed hold a fundamental sentiment that, for example, "whites" are higher status than"blacks", it would only make sense that Americans acted in ways to reaffirm this status relationship. Further, ACT supports the notion that settings change our perceptions, thus explaining to a certain extent how our behaviors towards members of the opposite race can differ in particular scenarios. However, the above factors indicate ways in which the ACT explanation of the paradox of race may be improved via LCSS' inclusion of a cognitive dimension.

First, LCSS can explain how Americans can claim to hold a prejudice that is distinct from a fundamental sentiment without removing the assumption that the fundamental sentiment drives behavior. LCSS argues that the form of the query "What do Americans think of African Americans?" is likely better geared towards accessing fundamental sentiments, while "What do I think of African Americans?" may illicit cognitive or institutional constraints that inhibit the fundamental sentiment from being expressed (Fiske et al., 2002; Lizardo and Strand, 2010; Srivastava and Banaji, 2011). Second, and in a related vein, LCSS allows us to consider the way in which individuals may be differentially influenced by and retain partial views of institutions and their representative knowledge structures. Finally, LCSS allows us to consider the notion that racial prejudice may be enacted through non-racial identities and biases towards particular cultural forms.

While I will consider this line of thought much more deeply in my dissertation, it suffices to say that two questions are of interest in this case study. First, it is important to discern how much LCSS really adds to ACT's explanation of the paradox of race. Are cognitive factors useful beyond explaining this curated list of factors I give above? This decision can be made empirically by testing the extent to which a statistical model based solely on ACT can predict individuals' prejudices in contrast with a model based on LCSS. If the LCSS model performs better on real data, it is at least some indication that the cognitive assumptions expressed by LCSS are useful beyond the ACT assumptions they are built on top of.

Second, and more generally, how can we use LCSS and the tools I develop to understand the paradox of race? What is the prevailing, general sentiment of Americans towards African American identity, and to what extent to different subcultures (with respect to user attributes) and differing forms of this identity (e.g. African American, black, etc.; Hall et al., 2015) mediate these sentiments? If negative prejudices exist for these identities, do individuals "hide" these prejudices in the language or acceptable actions of institutions? If negative prejudices do *not* exist, then where are the racial prejudices held that encourage existent racial inequalities? I have here touched on at least five possibilities where such prejudices could be held. Racial prejudice may exist within settings, contextual stimuli, the negative prejudices of entities the individual relates strongly to race, the differing extent to which individuals use particular institutions, subconscious biases or simply in institutions themselves.

In principle, the statistical model described in Chapter 5 should be able to inform me of the

extent to which prejudice exists within each of these elements, save for intricate subconscious biases that are not necessarily evident in text. Of course, the data used and the intricacy of the estimation process implies that the conclusions made on this case study will be preliminary and useful for further, more careful analysis rather than any conclusive explanation of the paradox of race.

7.1 Data

Approximately a week after the Michael Brown tragedy, I collected data for approximately two weeks using a keyword search on the Twitter Streaming API with terms relevant to the situation ("ferguson", "michael brown", "darren wilson"). I then stopped collecting data until the grand jury investigation finished, at which point I used several connections to the streaming API to search another set of relevant terms for approximately a week (#fergusondecision, McCulloch, brown, ferguson, michael brown, berkeley, darren wilson, police). Following the Eric Garner grand jury conclusion, I used the keywords "garner" and "eric garner" to search the Streaming API, again for approximately a week. I also developed a different keyword set to continuously capture data from the Streaming API around that time (icantbreathe , i cant breathe, garner, michael brown, freedom plaza, blacklivesmatter, alllivesmatter), which I ran for around five weeks.

On December 24th, 2014, I began using the Twitter REST API to extract the full set of tweets for all users that had sent at least five tweets in the full set of collected data. In total, there were just over 1.2M such users. I also extracted their follower and following relationships. Due to other uses of the machine on which this processing was occurring, collection took approximately five weeks. There are two restrictions to this collection process. First, Twitter only allows you to download historical data from users who are still active and those whose accounts are still public. Second, Twitter only allows you to see the previous 3200 tweets from any given user. Thus, for users like news agencies, who tweet frequently, I was not able to obtain all of their historical tweets. Regardless, I have the full, or nearly the full, timelines for approximately 1.2M users who can be expected to have tweeted actively about the Michael Brown and Eric Garner tragedies, as well as their follower and following relationships.

7.2 Statement of Work

As with the prior case study, I expect that research questions will evolve as I read further and via discussions with domain experts. In addition to/conjunction with exploring the questions posed in the introduction to this chapter (or modifications of them, pending further discussion), the following will necessarily be completed:

1. An "identity survey" and an "institutional survey" on the data, using the methodology described in Chapter 4
2. An estimation of the statistical model in Chapter 5 using this data
3. A general comparison of the semantic relationships of race as perceived by individuals with different demographics and by distinct institutions learned from the data
4. An analysis of the correlation between the follower/followee relationships in the data and similarities in perceptions overall and, in particular, of racial identities

Chapter 8: Conclusion

I have for much of my life been interested in how to reduce negative prejudices and the behavioral effects of these prejudices of others that is, how to lessen the negative perceptions individuals, groups, organizations and institutions carry of others. This problem concerns not only intergroup relationships, as it is often posed, but also has impacts on social issues like bullying, social isolation and mental health. Additionally, this problem concerns not only the human mind, as we often like to think, but also is encoded in the complex social structures within which we are ultimately embedded (Heise, 2007; Saperstein et al., 2013). Unfortunately, as recent work has suggested (Paluck, 2012), we need to better understand prejudice, and how it exists within the human mind and within these sociocultural structures, before real progress on prejudice reduction can be made. This thesis introduces new theory and new methods that I hope will serve to further our theoretical and empirical understandings of prejudice. In this section, I quickly review the contributions and limitations of the work as currently proposed. I then conclude with a timeline outlining the expected course over which the work proposed will be completed.

8.1 Contributions

8.1.1 Theoretical Contribution

LCSS provides a rich understanding of prejudice by extending on ideas primarily from Affect Control Theory, but also from schema theory, activation theory and strong practice theory. Importantly, it does so at a negligible increase in the conceptual complexity of ACT by working to synthesize a cognitive model into current ACT assumptions. LCSS thus attempts to fill gap and forms bridges rather than attempting to create a whole new conceptual structure. The primary gaps and bridges that LCSS constructs are the following:

- LCSS fills a gap in ACT where individuals are not able to think associatively, instead relying on institutions to form associative connections between entities. This allows LCSS a new interpretation of deflection, where individuals seek to maintain both semantic and emotional fundamental meanings. It should help LCSS to provide better predictions within social situations as compared to ACT. It also allows us to consider how emotional perceptions of identities may "filter" to other, similar identities
- LCSS fills a gap in ACT where individuals omnisciently infer institutional constraints from situations. It uses schema and activation theory to better describe how individuals infer appropriate institutional constructs via stimuli from social contexts.
- LCSS creates a bridge between cultural sociology and social psychology in unifying perceptions of cultural forms and identities under a single cognitive framework. This bridge is useful to the study of both culture and prejudice. With respect to the study of prejudice, it

helps to explain how prejudices of a group may be linked to skills or values associated with the group, rather than simply a view of the group itself. For example, some people may not dislike Republicans, but rather dislike people who do not believe in global warming, a large collection of whom are Republican. With respect to the study of culture, the affective measurement model of ACT brings structure and interpretive power to the content of perceptions of cultural forms. This measurement model is useful for an area of scholarship that has, at times, shown little regard for the affective content of one's perception of cultural forms relative to the perceived semantic relationships between them (Goldberg, 2011).

8.1.2 Methodological Contributions

Although I develop two novel tasks and at least one new model, this thesis will largely focus on the use of existing statistical and computational tools, rather than presenting any new algorithms or theory that will be applicable to larger classes of problems. The methodological contributions here thus serve chiefly to provide answers to new questions of interest to prejudice research using state-of-the art techniques.

Chapter 4 makes three major contributions:

- The creation of a novel and practically interesting NLP problem with corresponding gold-standard data from two different text modalities
- A model that will be of significant use to ACT scholars in moving from raw text data to the types of data that have been used to analyze the mathematical model of the theory in the past. Once individuals, identities, behaviors, settings and the events in which they are situated have been extracted, scholars can apply the theory directly if they so choose.
- The extraction of LCSS entities allows for an "identity survey" (Heise and MacKinnon, 2010) for these different datasets, as well as a way to frame the set of institutions that are relevant in the discourse.

Chapter 5 provides the following extensions over my prior work on extracting EPA profiles of identities and behaviors from text, and thus provides the following contributions to the literature:

- My prior work did not represent a schematic model of cognition, implementing only a model of ACT. Here, I propose a full model of perception in the form of an LCSS
- The prior model did not incorporate user attributes or even individual perceptions at all, only assuming that there existed multiple latent EPA profiles across the entire population of news articles. The proposed model provides a significantly more useful inferential tool by relating specific attributes to changes in perception
- The present model explicitly incorporates institutional knowledge. This will improve the predictive ability of the model, but also better represent the assumption that institutions influence the way in which people express their latent perceptions
- The previous model was static. By including dynamism in the model via a Markov representation of individual perceptions, we can gain an understanding of how perceptions change over time
- Evaluation of my prior work did not include rigorous simulation analysis[1] to ensure that the inference algorithm was performing correctly

[1] though I certainly tested the model with simulated data

- The proposed model will need to deal with significantly larger datasets and thus be much more efficient in its inference mechanisms than the prior work
- The proposed work includes a visualization component, which will mean that the results I develop will be more easily interpretable than output from the previous work

Beyond extending my previous work, the model I propose is the first I am aware of to attempt to develop and learn direct statistical model based on a theory of prejudice from raw text data. While others have used sentiment analysis or subspace analysis to get at prejudice indirectly, I believe that the model proposed here will have significantly more explanatory power.

8.1.3 Case Study Contributions

One primary use of the case studies in this thesis is to show the applicability of the developed theory and methods to two diverse scenarios. In one case, prejudices and identity relationships as perceived by individuals and news agencies are muddled. In the other, the prejudices and relationships of interest, as well as their form, are quite clear. These two scenarios thus present two unique vantage points from which to analyze the usefulness of the theory and methods presented.

In the case of the Arab Spring, this thesis will provide new tools to explain and to empirically explore the complex sociocognitive relationships in the region. These tools will provide new evidence that will lead to newer, stronger hypotheses on how these relationships fueled or abetted revolutions and violence. With respect to the Ferguson data, I am hopeful that the result of my theory and model are a new socio-cognitive explanation, with empirical backing, of the "paradox of racism" in America today. This work, if it succeeds, will be an important stepping stone to better understanding why current strategies for racial prejudice reduction so often fail, and may provide new evidence for alternative strategies that have a better chance of succeeding.

8.2 Limitations

8.2.1 Limitations of LCSS

There are several points which the theory of LCSS as currently described struggles with:

- As it stands, there is really no representation or even a discussion of how to think about the self in this model. LCSS' "self" would seem to fall somewhere in the middle of the self described by Miles (2014), the self described by Heise and MacKinnon (2010) and the original, "non-existent" self of ACT. Existing somewhere in the middle is a boon, and not a benefit, to the theory, as without a more comprehensive treatment of the subject, any use of the term "self" is unspecified and thus meaningless.
- Another thing I have struggled with immensely is how to think about the dichotomy of the queries "What do I think of people of type X?" versus the query "What do I think people of type X think of people of type Y"? As discussed by Rogers et al. (2013), the Stereotype Content Model (Fiske et al., 2002) generally assumes these to be distinct, while ACT does not make such a distinction. For the Ferguson case study, LCSS must develop a mathematical model of this distinction.
- The EPA position of cultural forms should be further considered. While I believe the mapping should be fairly straightforward, I currently know of little work (though I'm sure it exists) to justify this mapping.

While I propose several ideas on how to test LCSS throughout this document, the most important problem with the theoretical work is that I do not provide any concrete hypotheses as of yet. This means that there is no explicit test of specific portions of the theory, only use of the theory in developing the statistical models. I am working to remedy this issue, but will have to ensure that I develop a reasonably compact set of hypotheses in order to bound the scope of the thesis work. This is something I plan to fix before the proposal.

8.2.2 Methodological Limitations

The work in Chapter 4 suffers from at least the following two limitations:

- Entity extraction is independent of entity disambiguation, entity linking and coreference resolution, which are each considered possible post-processing steps. While surface forms dictionaries will implicitly account for some level of disambiguation, a model that can incorporate these processes into extraction may be more accurate (Moro et al., 2014).
- As I plan to use an off-the-shelf classification tool, my model will suffer where such an approach cannot capture inherent properties of the problem of study

The work in Chapter 5 suffers from at least the following five limitations:

- The model as specified makes strong parametric assumptions about the known number of contexts and institutions, which could feasibly be removed using a nonparametric approach
- The model makes strong functional assumptions (e.g. that we know the relationship between activation and likelihood a schema is emitted), which could feasibly be relaxed in a way where these relationships are learned from data
- The model makes strong distributional assumptions (e.g. the "schema space" is Gaussian) that could also be inappropriate
- The model relies on a "ground-truth" input of entity relationships and sentiments, which are themselves noisy processes with varying means of implementation.

8.2.3 Case Study Limitations

The most important limitations of my case studies are that I am not an expert in either the Middle East or race relations in the United States. Because of this, I will need to be careful to ensure that findings I uncover are appropriately scoped within the general knowledge base on these two domains. My only remedy for this issue is to constantly engage with scholars who are themselves experts in these areas and to better understand what empirical questions they have beyond those specified above that I may be able to answer using the proposed methodologies. Equally as important, of course, is that the data that I use creates important biases that should be considered at every stage of the analysis (see, e.g., Boyd and Crawford, 2012; Morstatter et al., 2013; Ruths and Pfeffer, 2014; Tufekci, 2014). While unavoidable, these issues will not be taken lightly .

8.3 Timeline

Figure 8.1 presents a rough timeline for this thesis. The timeline is rough in that I do not necessarily expect to "finish" portions of the thesis in their entirety- the links between the methodological and theoretical components of the theory are strong enough that I expect that insights gained when implementing the different portions of the proposed work will lead me to improve the other parts. However, I do expect to come to a more concrete formalization of the theoretical

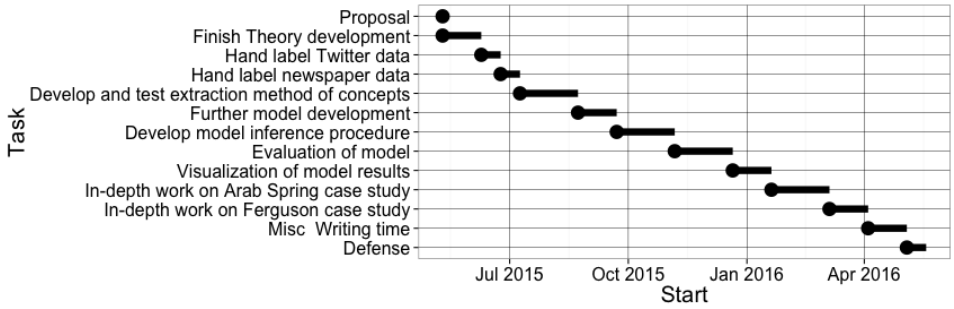


Figure 8.1: Proposed timeline of dissertation research

component first, and then modify it as necessary as I continue to read, engage in discussion with other scholars, and begin implementation of the two methodological components of my work.

With respect to these components, it will be necessary to complete, or at least mostly complete, the work in Chapter 4 before I begin working heavily on Chapter 5. I would also like to complete the hand-coding of the gold-standard data as soon as possible, both to further dig into the strengths and weaknesses of the data in my case studies and to be able to quickly develop at least simple models to address the problems stated in Chapter 4.

It seems feasible that all portions of this thesis can be completed in approximately fifteen months of work, leaving me to defend my thesis in late July of 2016.

Appendix A: Literature Review

This thesis relies heavily on literature and/or methods in several different fields, including cultural sociology, social psychology, cognitive psychology, natural language processing and statistics. Although this is the "larger" literature review in this proposal, it still is written in a compact form so as to emphasize the points most relevant to the proposal. I am in the process of synthesizing additional prose that I have written which explains the concepts here in even further detail. This work will be integrated into the dissertation. This appendix is split into two separate sections, one concerning the theoretical portion and one the methodological portion of the thesis.

A.1 Theoretical model

A.1.1 An overview of ACT

I begin with a cursory overview of ACT, fully acknowledging that in the space allotted, any attempt made cannot do it full justice. As such, I point the reader to Heise's (2007) book for a full treatment of both the conceptual and mathematical components of the theory, or to Robinson et al.'s (2006) chapter-length overview for a slightly shorter yet still reasonably comprehensive overview.

As suggested by Heise and MacKinnon (2010), ACT consists of three main assumptions, or components. First, it assumes a particular measurement system for identities, behaviors, settings and modifiers, which I will collectively refer to as *entities*. The measurement system induces a connection between linguistic categories that represent the names of entities and stable, affective meanings of these entities. Each entity is thus assumed to be represented by a particular linguistic form that has a specific, *fundamental* affective meaning that is shared across large cultural groups (e.g. national cultures). The fundamental affective meaning of each entity is defined by a sentiment located in a three-dimensional affective space, conceived of and validated empirically by Osgood (1969).

This three dimensional space is defined by axes entitled Evaluation, Potency and Activity (EPA), each spanning the range of -4.3 to +4.3. The evaluation dimension describes the goodness/badness of an entity. The potency dimension describes the powerfulness/weakness of a given entity, and the activity dimension describes its level of activity/passivity. 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. 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. The methodology involved in these surveys has evolved over time, and a thorough discussion can be found

in (Heise, 2010b).

As detailed by Heise and MacKinnon (2010), the link between linguistic categories and stable affective meanings of these categories develops over time as a particular culture's "theory of people" develops both the categories that should exist and the affective meanings of them. These meanings are imbibed by new members of the culture via socialization, and although they may change, these changes are relatively gradual and reflect shifts in culture-wide shifts in sentiment. ACT's emphasis on the relationship between linguistic categorizations of entities and relatively stable affective meanings makes the theory a particularly attractive candidate for use in natural language processing.

While I will shortly find reason to define the other entity types, it suffices here to introduce only the meaning of an identity in ACT. Smith-Lovin (2007) defines identities as the ways in which an individual can label another individual with whom they have had an interaction. She continues to defined three general types of identities. *Role* identities indicate positions in a social structure (e.g. doctor). *Category* memberships come from "identification with some characteristic, trait or attribute". *Social* identities indicate membership in social groups. As defined by Tajfel and Turner (1979), a *social group* is 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. This definition relies on the definition of a *social category*, which Cikara and Van Bavel (2014) define as "inclusive [social] structures that require merely that all members share some feature" (e.g. gender, race). As Jonathan Morgan is considering, while role identities create a duality between identity and the institution, social (category) identities thus create a duality between identity and the social group (category).

The connection between identity, social group and social category is important in defining the relationship between perceptions of identities, as described by ACT, and the meaning of prejudice in the psychological social psychology literature. As defined by Hewstone et al. (2002), a *prejudice* is a biased attitude towards a social group or category[1] The affective perceptions of identities defined by ACT thus *are* prejudices. Indeed, EPA profiles of ACT identities show strong correlations with stereotypes and prejudices of social groups as defined by the Stereotype Content Model (SCM; Fiske et al., 2002) and the SCM's associated affective component (the BIAS map; Cuddy et al., 2007) (Rogers et al., 2013). The SCM and the BIAS map are more heavily focused on stereotyping and prejudices than ACT and derive from a different lineage of literature, thus suggesting the usefulness of the latent space approach in conceptualizing a parsimonious representation of prejudice in cognition.

The second main component of Affect Control Theory is an empirical framework for how social events change our perceptions of entities. A *social event* is a social interaction in which an *actor* enacts a behavior on an *object*, perhaps within a particular setting (Heise, 2007, pg. 36). In most cases, both the actor and object are represented as instantiations of particular identities. ACT's empirical framework defines how "pre-event" EPA ratings of the actor, behavior, object and setting change after observing the event. The difference between the pre and post-event impressions of an individual who observes (or engages in) a social event is calculated using a

[1]Although not considered here, it is useful to contrast this definition with the definition of a *stereotype*, which is simply the association an individual perceives between an individual and a social category or between two social categories (Hewstone et al., 2002). The distinction between these two concepts is relatively muddled - though I consider this further later in the thesis, I will treat the terms as referring to the same general concept here.

change equation, which 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. Change equations are estimated empirically using survey data via a variety of regression techniques that have evolved over time (Heise, 2007; Smith-Lovin, 1987).

The final assumption of ACT is a cybernetic control system which specifies how actors will behave in particular situations. ACT is a "control theory" in that it assumes humans seek to maintain fundamental meanings of identities and behaviors in the transient impressions that are generated when social events are observed or carried out. While we may try to maintain these meanings through various methods (see Heise, 2007), our efforts are all carried out in an attempt to reduce the *deflection*, or difference, between our perceptions prior to and after the event occurs. ACT assumes that events which 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[2]. Importantly, the setting in which the event occurs may also have a strong impact on transient impressions, and thus our prejudices may be impacted by the place and times at which social events occur Smith-Lovin (1987).

Along these lines, an important distinction is made in ACT between "setting" and "situation". A setting, as noted above, is a linguistic category that define a place and time with a specific fundamental meaning. Situations, on the other hand, are better described as the entire social process in which interactions take place. One important part of a social situation is what I will define as the *context* of the situation, which provides the cognitive stimuli that individuals use to define their perception of themselves and others around them in a particular situation. In ACT, context provides the cue for which *institution* is appropriate in a particular situation. Heise (2007) defines an *institution* as a "constellation of identities, settings, and actions relating to some elementary concern" (pg. 28). Institutions "organize the huge number of identities that you can encounter" and constrain the possible identities we take on and behaviors we are willing to enact. While certain identities exist outside of institutions friend, for example- most identities make sense only within institutional boundaries. Thus, ACT uses the concept of institutions to help explain why someone might take on the identity of a doctor in one context versus an engineer in another, where both may have appropriate EPA profiles useful for rectifying the social events occurring during any given social situation.

A.1.2 The Mathematics of ACT

Equation (A.1) gives a nine dimensional vector that contains the EPA profiles associated with the three entities ($a$ctor, $b$ehavior, $o$bject) in a social event, where I here do not consider additional elements such as settings or modifiers.

$$f = \begin{bmatrix} a_e & a_p & a_a & b_e & b_p & b_a & o_e & o_p & o_a \end{bmatrix} \tag{A.1}$$

Given the form of the fundamental, we can now describe how a social event changes these sentiments to produce a post-event transient impression. This change occurs via the application

[2]these values were computing using the INTERACT Java program (Heise, 2010a)

of the change equation to a particular fundamental vector f. Though a variety of estimation methodologies have been utilized to estimate the change equation (Heise, 2007), the form of the equation is expected to define a linear combination of the fundamentals. 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:

$$\begin{aligned} G(f) =[1 \quad & a_e \quad a_p \quad a_a \quad b_e \quad b_p \quad b_a \quad o_e \quad o_p \quad o_a \quad a_e b_e \quad a_e o_p \quad a_p b_p \quad a_a b_a \\ & b_e o_e \quad b_e o_p \quad b_p o_e \quad b_p o_p \quad a_e b_e o_e \quad a_e b_e o_p] \end{aligned} \tag{A.2}$$

Second, for each element of $G(f)$, we can define a set of coefficients, M, that describes the extent to which the element modifies ethe 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.

As noted above, the deflection of a particular event is a measure of the difference between the pre- and post-event impressions. As we are always assuming that the pre-event impression is set equal to the fundamental, we can define deflection as the squared difference between the fundamental and the transient impression resulting from the event, as shown in Equation (A.3). In the equation M_{i*} is the ith row of M.

$$Deflection = \sum_{i}^{9} (f_i - M_{i*}^T G(f))^2 \tag{A.3}$$

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. While one could, quite simply, provide a pathological 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 pathological case. 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.

A.1.3 Limitations of ACT

By combining a measurement system for entities, a mathematical model of how social events change transient impressions of entities and a cybernetic control assumption that defines the ways in which people can be expected to react to these changes, ACT presents a powerful mathematical model for understanding impressions of identities and how these impressions can temporarily change in social situations. Using ACT, we can, among many other things, predict expected behavior of one identity on another, predict the prejudice an individual taking on one identity has of another individual taking on a different identity given the behavior she chooses to enact and understand the effect of particular contexts on prejudices of of particular identities. However, there are several limitations to the theory. In particular, there are at least five ways in which to the theory as it currently stands could be extended to be even more useful. While the first two of these have already been addressed, the final three limitations are still, to the best of my knowledge, open for exploration, and are the focal point of the theoretical work in this thesis.

First, prior to 2010, ACT did not have a published, formal theory of the self. Heise and MacKinnon's (2010) book developed a new theory of self that relies on similar cybernetic assumptions to ACT. More specifically, Heise and MacKinnon (2010) assume the existence of a *self-sentiment*, an individual's sentiment of herself as a singular entity, or "persona". They rigorously define the self, how perceptions of one's self influences behavior (Heise and MacKinnon, 2010, see the diagram on pg. 126), how social interactions can, over time, change one's self sentiment and how these processes exist under a cybernetic control system and in connection with ACT. In doing so, Heise and MacKinnon (2010) provide a vital tool in understanding how recurrent behaviors, contexts and identities taken on by individuals can have long-term effects on future actions, a tool that did not exist formally before their efforts.

More recently, Miles's (2014) developed a theory of identity and culture a model which drew heavily on Heise and MacKinnon (2010). Miles's (2014) work provides a pragmatic tool for understanding how we determine identity in a particular social situation. Perhaps most importantly, Miles continues a line of argument from Identity Control theorists who describe identities as cognitive "schemas". Over time, a set schema (identities) which become perpetually accessible and come to define the "core self". Only when we strongly feel our core self does not fit in a particular context to we activate other identities in our identity hierarchy which we find more appropriate. While I, like Heise and MacKinnon (2010), have reservations over the idea of a hierarchy of identities which are transmittable across situations, I draw on the schema-theoretic concepts introduced by Miles's (2014) in developing my model of self in LCSS.

The second limitation of ACT is that the mathematical model is unable to incorporate uncertainty individuals have in their own identity, the identity of other interactants, and the EPA positions of these identities within social situations. Similarly, it is unable to express mathematically the varying extent to which fundamental meanings may change over long periods of time. Recently, 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, in doing so, Hoey and his colleagues introduced *Bayesian ACT (BayesACT)*, a mathematical model that was able to incorporate uncertainty into theory predictions. Additionally, through tuning of priors on the variance of fundamental meanings, BayesACT provides a concrete means of expressing how much one expects fundamental meanings to change over time. More recently, Hoey and Schröder furthered this model to incorporate the self-concept introduced by Heise and MacKinnon (2010) into their model. I draw heavily on the mathematics of BayesACT in LCSS and have utilized them in prior work inferring EPA profiles from newspaper data (as discussed in Appendix B).

The third limitation of ACT, and one inherent in the theory as I understand it to exist currently, is that ACT is what I term a "first-order social effects model". ACT models the transient meanings of individual interactants taking on particular identities and how these effect each other, recursively, with continued interaction in a social situation. However, it does not define what I consider to be the "second-order social effects" of how social events between particular entities may impact perceptions of related entities. Indeed, such second order social effects are not possible to model with ACT, as the theory does not allow for cognitive internalization of existent relationships between identities (e.g. that "patient" is related to "doctor"). Thus, it cannot allow for prejudices of related concepts to "filter" to each other (e.g. prejudices of "men" to filter to

prejudices of the unfortunately predominantly male occupation of "engineer").

This inability to model second-order effects is important because the human mind is likely to make these sorts of inferences. Schema theory (Rumelhart, 1978) provides one model of cognition that explicitly shows how humans make these sorts of associations. Schema theory posits that associative memory can be modeled via *schemas*, which are "data structures" that hold sets of variables which define what we expect from a particular concept Rumelhart (1978). Importantly, schema do not exist in isolation. Rather, schemas are situated within networks, where links connect schema that our brains have come to believe are related to each other. While the general use of schema theory with relation to identities is not new (Miles, 2014; Stryker), the emphasis in this thesis is in how the explicit mathematical modeling of these schema within a network can better inform our understanding of how and why people adopt particular prejudices of these identities.

To deal with the lack of an explicit model of cognition, ACT introduces institutions as overarching "knowledge structures" that inform individuals of semantic relationships between identities. A fourth limitation of ACT, I argue, is that this definition of the institution is ambiguous and in being so appears to ascribe too much power to institutions to drive interpretation of context. Such ambiguity becomes clear in comparison to discussions of the institution in the cultural sociology literature. The use of literature on culture is relevant because prejudices themselves meet contemporary definitions of culture (e.g. DiMaggio, 1997) and more cognitive cultural elements can still be used as the bases for affective prejudices (Cikara and Fiske, 2013). Thus, studies of culture are inclusive of the study of prejudice, at least at a high level.

Until the later portion of the 20th century, the dominant view of culture in sociology was that it was a coherent *"system of beliefs"* encoded in the mind of individuals. Humans, it was argued, learned this system of beliefs as children from, for example "parents, schools, and churches" (pg. 1577 Vaisey, 2009). Individuals were then expected to use this system of beliefs, encoded in institutional doctrine to motivate their actions. The downfall of this perspective is best associated with the work of Swidler (1986), who noted that individuals were terrible at understanding why they acted in particular ways in most situations Vaisey (2009). Swidler's (1986) work thus suggested that while systems of beliefs may exist at the institutional level in doctrine, these systems were not responsible for motivating social action.

Swidlers "toolkit" theory instead starts with the assumption that humans are cognitive misers (Simon, 1987), internalizing only a small set of habits, skills, heuristics and practices that we can almost always rely on to inform our actions. In the case that we cannot find a heuristic to match a specific situation, we fall back on "publicly available" (pg. 208 Lizardo and Strand, 2010) systematic knowledge to inform us. Toolkit theory thus assumes that people are not "socialized" in the traditional sense of being taught how to act, both because we cannot hold an internal model of the complex, systematic world around us and because the process of "socialization" itself is not what instills within us the heuristic knowledge we use to inform our actions on a daily basis.

While toolkit theory moved towards a more cognitively plausible model of culture, Lizardo and Strand (2010) suggest there exists a middle ground between the "old" sociological model of systemic culture and toolkit theory that may be preferable. This middle ground is referred to as *strong practice theory (SPT)* . Strong practice theory, which draws heavily from the work of Bourdieu (1986), agrees with toolkit theory in that it considers culture to be more a set of heuristics that we use to inform our actions than some kind of knowledge "system" that we

acquire through the socialization process. However, strong practice theory departs from toolkit theory in that it believes individuals are capable of internalizing a rich and complex system of these practices and habits, rather than only internalizing a crude and shallow set of rules that we consistently apply.

Drawing on dual-process models of culture in action (Vaisey, 2009), Lizardo and Strand (2010) argue that this internalization exists at the level of associative memory; in other words, that culture is systematic at the level of cognitive schemas . The embedding of a complex set of heuristics and practices is what allows for regularity in human behavior, to the extent that there appears to be a rigidly internalized "system of beliefs" in many cases. On the other hand, how we interpret and justify our own actions is a function carried out at the level of discursive consciousness. Our post-hoc interpretations of (our own!) heuristics that are too deeply internalized in associative memory for us to explicate coherently are, Lizardo and Strand suggest, what produce the inconsistencies that led to the development of toolkit theory. Consistent with toolkit theory, in this process of justifying the actions we take, we rely not only on an interpretation of our own heuristics but also on the "cultural scaffolding" available to us from social institutions. Strong practice theory thus suggests that while automatic processes that are deeply internalized within schematic networks often inform our actions, the way we explicate these behaviors is an unsteady inferential procedure that synthesizes what we can capture from our own complex cultural knowledge and the scripts provided to us to explain actions by the institutional doctrine around us. Thus, it may be less that our perceptions are driven by institutions, and rather that the way we justify or come to terms with our own actions and those of others are what is effected (Srivastava and Banaji, 2011).

In this way, strong practice theory outlines the fact that culture exists, in large part, at the intersection between the information that is available to us in the social environment around us and the way we internalize this culture in our cognition DiMaggio (1997). Institutions thus often act more as tools for explaining actions from underlying cultural skills, habits and knowledge than as mechanisms that produce or affect this culture. Srivastava and Banaji (2011) propose a similar dual-process approach to understand how people internalize and use institutional knowledge to justify their behavior. They suggest that associative (automatic) consciousness frequently drives behavior regardless of institutions, but that these behaviors are justified using cues from institutions.

Although (Heise, 2007, pg. 71) does refer to "enculturation", its Symbolic Interactionist (Mead, 1925) roots suggest ACT is not a pure "system of beliefs" model. However, it is difficult to distinguish ACT as either a toolkit or a strong practice model in its proscription of the role of the institution because ACT explicitly chooses not to explain how identities are selected (Heise and MacKinnon, 2010, pg. 200). Because of this, there is no explanation of how individuals "know" what identities are appropriate in what institutions" (Heise and MacKinnon, 2010, pg. 209). In other words, the theory does not explain how, and is ambiguous as to if, individuals internalize a culture's "theory of people" that is represented within institutions. LCSS aligns itself strongly with the SPT perspective on cognition and institutions, arguing that much of what is encoded in institutional knowledge structure is also encoded in our (sub)conscious minds. Only when these two representations differ do institutional knowledge structures play a role in our understandings (Lizardo and Strand, 2010). Because it does not try to model cognition, ACT also does not define how contexts inform actors of which "institution's cues predominate at a

given time and place". LCSS uses *activation theory* Anderson et al. (1997) to detail this process of how situations inform actors of available institutions.

Activation theory suggests that contexts provide stimuli which in turn induce schemas related to particular entities. These entities may themselves inform us of what institutional knowledge is applicative in a particular social situation. These relationships between contexts, entities and institutions may be, and are likely to be, stable, however, they are by no means given or omnipotent. Like fundamental meanings, these connections must also be learned. More specifically, activation theory holds that stimuli *activate* schemas the mind has learned to associate with these them Anderson et al. (1997). Schemas that are activated by the same stimuli are, over time, see as being related to each other. Additionally, activation spreads from originating schema to other related schema through the process of spreading activation (Collins and Loftus, 1975). Thus, stimuli may in time become associated with collections of schema through repeated simultaneous activation, thus leading to "institutional perceptions" within the mind. When a schema reaches a level of activation beyond a particular threshold, it is instantiated. Our mind uses the set of instantiated schemas to "fill in" missing information in the current environment, providing us with a more informed sense of what we should expect from our current situation. Where these schemas are identities, they may be used to frame our perspective of the self and/or others (Miles, 2014).

The final way in which I believe ACT can be extended is by including perceptions of *cultural forms* into the model. Of course, the term cultural form is generic, and could be seen as encompassing personal "traits" or "attributes" Heise and MacKinnon (2010) or more generally to "include the knowledge and skills highlighted [and] motivating constructs like attitudes, values, and moral worldviews" (Miles, 2014, pg. 212). Further, definitions of the term culture are themselves incredibly varied - even in 1952, there were over 150 definitions of the term (Kroeber and Kluckhohn, 1952). I restrict my work here to focusing on cultural forms as defined largely by cognitive elements like skills, values and preferences, drawing from the use of culture as such in Lizardo; Lizardo; Lizardo's (2006; 2011; 2014) line of work on culture in action. From a cultural studies perspective, ACT provides an important model for the *content* of cultural perceptions in a field that is primarily concerned with relationships, or associations, between cultural forms Goldberg (2011). More generally, the merger of theories of identity with the cultural sociology literature has already been shown to be a fruitful avenue for better understanding the *use* of culture Lizardo (2006, 2011), how cultural capital can affect an individual's ability to enact particular identities Bourdieu (1986) and the actions that individuals may take to confirm or disengage from an identity Bourdieu (1990). I build on this work by providing a more mathematical framework around these ideas, best and most recently compiled by Miles (2014).

A.2 Empirical Methodology

In this section I give an abridged discussion of the tools and techniques that are either utilized by or similar to the empirical methods I propose for my dissertation[3]. A more complete discussion will also be included in the final dissertation.

I begin with a discussion of the extraction of higher order concepts from words. While we

[3]For more in depth reviews of the use of text processing in the social sciences, I forward the reader to the excellent works from OConnor et al.'s (2011) and Grimmer and Stewart's (2013)

can gain much by interpreting language as a "bag of words", the theoretical ideals of LCSS rely on obtaining information on individuals, social categories and cultural forms. These objects may exist across multiple words and different collections of words may refer to the same object. Thus, we must give some thought to the use of concepts, instead of simple terms, for the purposes of the proposed analyses. I then move to a discussion of how to extract relationships between these concepts. As we will see, the extraction of relations between concepts naturally leads to a discussion of how to represent them in a latent space.

A.2.1 Entity Recognition, Extraction and Linking

The extraction of mentions of concepts from raw text is known broadly as *entity mention detection*. The most straightforward solution to this problem is to consider concepts as being represented by sets of n words that frequently occur together in text. These sets of words may be compounded into an *n-gram* and be used to represent a single concept. While the extraction and use of n-grams can improve results in many NLP tasks, extracted ngrams are not guaranteed to represent concepts of interest in the real world. One remedy for this is to use *part of speech tagging* (POS tagging), a technique that determines the part of speech of every term in a sentence. Part-of-speech taggers exists for a host of languages as well as for social media dialects like Twitter (Gimpel et al., 2011; Owoputi et al., 2013). With a part-of-speech tagger, one can choose to only extract n-grams (including single terms) that represent frequently occurring noun phrases, thus increasing the likelihood that the referred-to concept is actually of practical interest. To further increase the likelihood that surface forms extracted from text are representative of real-world objects of interest, a final popular approach to concept extraction is to use thesauri, or dictionaries, to map from linguistic representations (*surface forms*) of particular objects into a single higher-order concept (e.g. Wikilinks[4]; CrossWiki[5]).

Regardless of how entity mentions in raw are detected, however, one may then be concerned with determining with high probability those mentions that refer to the same real-world concept of interest. Tools to make this decision are generally concerned with one of two problems. First, one may be concerned with determining which real-world entity a surface form refers to (e.g. does "Michael Jordan" refer to the basketball player or the statistician?). This problem of *entity disambiguation* has seen both heuristic graph-based and more principled statistical solutions. A second, related problem is determining if two surface forms refer to the same higher-order concept (e.g. do "bike" and "bicycle" refer to the same thing?). As noted, this process of *coreference resolution* may be carried out via the use of a thesaurus which maps many surface forms into a single concept. Similarly, one may be interested in resolving pronoun usage within text. In this case, as with coreference resolution more generally, both deterministic Lee et al. (2013) and probabilistic approaches Soon et al. (2001) have been shown to work very well in practice.

After running entity mention and disambiguation, we have at our disposal a set of n-grams that refer to unique real-world concepts. A natural step at this point is to see if we can collect more information about this particular concept. The task of *named entity recognition* (NER) represents the extraction of concepts from text and their categorization into a general typology

[4] `http://www.iesl.cs.umass.edu/data/wiki-links`
[5] `http://www-nlp.stanford.edu/pubs/crosswikis-data.tar.bz2/`

of named entities, most often people, places and organizations. Tools providing NER allow us to extract a typology of named entities in our text, but cannot help with our understanding of more general concepts. This may help us to, for example, understand the higher-order topical focus of a particular text. Knowing that Sidney Crosby is a hockey player may help us to better understand that a tweet or newspaper article is about ice hockey (Han et al., 2014). The connection of concepts extracted from text to external information on the entity is known as the process of *entity linking* . Entity linking is performed from text into entities existing in a *knowledge base* Medelyan et al. (2013), a large data source providing (semi-)structured information on millions of real-world concepts. The canonical example of a knowledge base is Wikipedia, but other examples include Yago2 Hoffart et al. (2011, 2013), NELL (Mitchell et al., 2015a) , Freebase (Bollacker et al., 2008) and Wikidata (Vrandei and Krtzsch, 2014).

The concept extraction I wish to perform here is unique in two important ways. First, part of my analysis will be performed on Twitter data. While a significant amount of work on the above tasks has been focused on Twitter (Chang et al., 2014; de Oliveira et al., 2013; Ferragina and Scaiella, 2010; Gattani et al., 2013; Gimpel et al., 2011; Guo et al., 2013; Kong et al., 2014; Lin et al., 2012; Meij et al., 2012; Owoputi et al., 2013; Ritter et al., 2011; Wang et al., 2012; Yang et al., 2014), few of these efforts have resulted in toolkits that can perform the desired tasks have yet to be developed. Second, and more importantly, the entire workflow discussed above is not particularly interested in the extraction of general identities (e.g. "teacher"). Similarly, and particularly on Twitter, approaches that rely solely on knowledge bases for entity dictionaries are ill-suited to the extraction of novel entities that emerge during particular events Han et al. (2014). These issues thus suggest that, while the tools, methods and approaches may already exist to extract the concepts of interest to my theoretical model, they do not currently exist in any unified form. I will thus need to combine useful output from these previous works as features into a model that can be used to extract the identities, cultural forms and individuals of interest.

A.2.2 Extracting relationships between concepts

The general problem of extracting relationships between concepts within text is massive, encompassing a variety of tasks and approaches. Here, I concentrate on two types of relationships between concepts, co-occurance and linguistic dependencies, and two general problems, inferring the type and meaning of connection between concepts and inferring higher order, topical relationships between concepts.

Most relevant to the discussion above is the process of *relation extraction*, which uses statistical regularities across many documents to extract typed relationships between concepts (e.g. "[Obama] [born_in] [Hawaii]"). Example systems include REVERB Fader et al. (2011) and NELL Mitchell et al. (2015a), which crawl the web to extract such relations between concepts and to store those relations into knowledge bases. Most often, these approaches utilize some form of dependency parsing Kbler et al. (2009), which uses statistical models to determine linguistic dependencies between concepts within a particular text. For example, from the sentence "The teacher advised the student" a dependency parser (and post-parsing lemmatization) could extract the relationship "teacher advise student".

However, perhaps the best known method for determining relationships between concepts (or more frequently, just words) in text is to simply use the co-occurence of the concepts in the same textual unit (e.g. a sentence, paragraph or tweet). Direct utilization of these co-occurance

relationship is the basis for semantic network analysis Carley (1990); Carley and Kaufer (1993), which is useful in understanding the latent "mental map" that authors of texts have between different concepts. Co-occurence relationships in text can also be used to extract higher-order topics from text, where terms that repeatedly occur in the same text unit across many documents are assumed to be drawn from the same topic.

Perhaps the best known approach to extracting topics from text are unsupervised admixture models, known generally as *topic models*. The canonical example of a topic model is Latent Dirichlet Allocation (LDA; Blei et al., 2003). Blei's LDA model assumes that each document can be represented as a distribution over latent topics, each of which is itself a distribution over all words in the document corpus. The model has been used in a variety of settings, from understanding large academic corpora to applications beyond text data, for example my own work in applying LDA to foursquare check-ins (Joseph et al., 2014b). Unfortunately, while topic models have seen a rapid increase in popularity in both the machine learning literature, where thousands of variants have been proposed, and also in the social sciences, with entire special issues devoted to simply understanding topic models and their applications Mohr and Bogdanov (2013), there has been a relatively low level of overlap between advances in social theory and advances in Machine Learning[6].

While many social scientists are only now beginning to come around to the power of unsupervised topical extraction, statistics and Machine Learning researchers have gone well beyond these original models. In addition to the basic extraction of independent topics, we can now learn correlated (Blei and Lafferty, 2007; Kim and Sudderth, 2011), hierarchical (Blei et al., 2004; Li et al., 2012b), and dynamic (Blei and Lafferty, 2006; Takahashi et al., 2012; Yin et al., 2011) topics in an unsupervised fashion. We can incorporate language modeling with topic models to jointly extract concepts and topics together (Han et al., 2014). We can also incorporate external information to develop semi-supervised (Ritter et al., 2011) or even fully supervised (Yang et al., 2014) approaches, or to learn additional information, for example topics specific to specific social communities (Liu et al., 2009) at specific points in time (Hu et al., 2013) or space (Sachan et al., 2014) or who have particular opinions on particular topics (Mukherjee, 2014). We can also compare the use of topics in documents with different features to better understand how topics correlate with predictors of interest (Mimno and McCallum, 2012; Rabinovich and Blei, 2014; Roberts et al., 2013). We know more about how different topic models function in data from different media (Hong and Davison, 2010), how humans interpret the output of topic models (Chang et al., 2009) and how we can evaluate the fit of a particular model to the data (Mimno and Blei, 2011; Wallach et al., 2009). Finally, a number of new estimation techniques have been proposed to make such models (and often, Bayesian inference more generally) faster and more scaleable (Arora et al., 2012; Bi et al., 2014; Neiswanger et al., 2013).

Note that topic modeling and models that infer typed dependencies between individuals are not unique. As a concrete example, O'Connor et al. (2013) 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. As my previous work has done, I will follow

[6]The Poetics special issue is a particularly exciting exception to this

in this vein, combining the nuances that typed dependencies between entities allow for while still utilizing the latent variable notions introduced by topic modeling.

A.2.3 Representing text in a latent space

Topic models are particularly useful in that they place text into a latent space with a constrained dimensionality. From this latent space, interpretation, in particular visualization (Dou et al., 2013), becomes much more straightforward. Such a space is precisely what the EPA profiles of ACT represent, and I will create and attempt to learn a similar embedding of concepts into a latent space using a statistical model. In addition to topic modeling, more recent approaches involving neural networks and "deep learning" (Lee et al., 2009) have come to the forefront of vector space representations of both words and concepts. These tools, trained using a variety of features extracted from many of the processes described above, have been shown to be useful in both understanding meaning and in prediction problems (Mikolov et al., 2013).

A.2.4 Sentiment Mining

At their root, EPA ratings of concepts are a theoretically validated and very rich source of sentimental meanings. The extraction of the sentimental meaning more generally of different terms in a text is far from novel in the NLP community ((e.g. for Twitter Calais Guerra et al., 2011; Gao et al., 2014; Guerra et al., 2014, 2013; Hutto and Gilbert, 2014; Li et al., 2012a; Liu et al., 2012; Taddy, 2013) and more generally (Cambria et al., 2014; Dong et al., 2014; Maas et al., 2011; Wang and Manning, 2012; Zhu et al., 2014, e.g.). 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); for its application to the social sciences, we refer the reader to Grimmer and Stewart (2013), Sections 5.1 and 5.2; and for a shorter but more updated review we refer the reader to Section 2.1 of Dong et al. (2014) or to Mukherjee and Bhattacharyya (2013).The work I propose for my dissertation differs from most work on sentiment analysis in three ways.

First, as opposed to estimating the sentiment associated with an entire text, we are much more interested in the sentiment associated with particular concepts. While work exists at this *concept-level* Cambria et al. (2014) (or *phrase-level* Dong et al. (2014)) of sentiment analysis, it is often developed with the purpose of finding words that are indicative of positive or negative global document sentiment. For example, Maas et al.'s (2011) develop a model that embeds words in a latent space which includes information about word sentiment from the document level, allowing them to find sentiment-discriminating terms. An exception to this focus is the work of Cambria et al. (2014), who develop a statistical model that incorporates both sentiment and a form of schema theory, considering the flow of "energy" between concepts in a cognitive model. However, their work emphasizes the existence of a single, global knowledge base to capture cultural and sentient meanings.

This captures the second distinction between my interest and other work. In contrast to most efforts, my theoretical model suggests there is no such global structure; rather, culture is distributed and only exists at a global level via institutional knowledge or when the researcher determines some method of aggregating over individual cognitive models. With respect to uncovering individual-level sentiments, work by Mukherjee's (2014) develops a joint model of authors

and their sentiments towards different topics. With respect to the principled aggregation of these opinions across individuals, Hoang et al. (2014) develop a model that learns groups of individuals who have a common sentiment towards particular topics. A similar concept is the induction of signed social relationships between actors using textKrishnan and Eisenstein 2014. These relationships represent sentiments, or prejudices, individuals have towards other individuals.

Finally, while many sentiment analysis approaches evaluate concepts on a single, evaluative dimension, our work places concepts into a richer, more descriptive three-dimensional latent space. These efforts are in a similar vein to recent work by Kim et al. (2012), who learn a multi-dimensional representation of concept-level sentiment scores.

Appendix B: Overview of Prior Methodology

This section gives a cursory description of my previous relevant work using text extracted from the full article, which is currently in submission. For more details, I refer the reader to a draft of the full article, available at `https://www.dropbox.com/s/fndqd1bfqrasp47/jms_submission_2015.pdf?dl=0`

B.1 Overview of approach

ACT defines a *social event* as a situation in which an *actor* enacts a *behavior* on an *object*. Further, the theory assumes that both the actor and the object are *identities*, which are nouns that define or allude to a social category, or an individual member of one, that we have a particular sentiment towards. My prior work developed a methodology that allowed for a better understanding of how cognitive constraints affect perceptions of a particular set of identities engaging in a particular set of behaviors across many social events.

Specifically, I, along with my colleagues, developed an approach that was able to infer latent affective meanings of identities and behaviors from the Arab Spring newspaper corpus. Our work was the first effort we were aware of to apply ACT concepts in an automated way to text data. 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. 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 (1969).

Two chief methodological issues were overcome when developing the model. First, while there are an increasing number of databases and tools for extracting world events (e.g. GDELT; Leetaru, 2011) 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 identity categories of interest (e.g. "teacher"). I used dependency parsing Kbler et al. (2009) to extract social events from a population of identities and behaviors of potential interest. However, I had to manually clean the resulting output to pull out interesting identities and behaviors from the noisy result of the dependency parse.

Second, given this set of identities and behaviors, I had to build a model to overcome ACT's methodological and data limitations. The primary contribution of my prior work was thus a Bayesian Network that provided an initial and substantial step forward in this direction. The model we introduced has four desirable features in that it:

- inferred affective meanings for identities and behaviors not currently in the ACT dictionaries
- provided a variance for the position of each sense of each identity and behavior
- inferred where multiple "senses" of a particular identity or behavior existed within the data
- incorporated prior knowledge from existing ACT dictionaries

In the prior work, we provided a formal evaluation of the model's ability to predict left out data and a case study of its feasibly. With respect to the former, we found that our model's performance improved over several baseline approaches on the prediction task, though it struggled with issues of data sparsity in comparison to the strongest of our baseline models. Still, the final model we presented gave meaningful semantic positions for the identities and behaviors within the dataset, which none of the baseline models were able to provide. With respect to the latter, we observed a discrepancy in the way in which major English-speaking news outlets portrayed the generic Muslim identity as opposed to the more specific Sunni and Islamist identities.

B.2 Overview of model

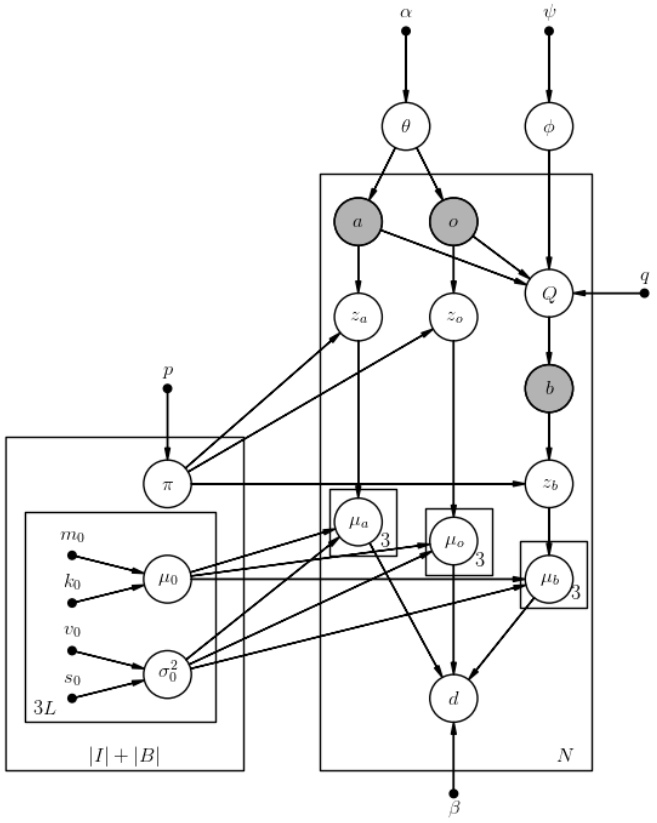


Figure B.1

Figure B.1 depicts the probabilistic graphical model utilized in my prior work. Although the

model appears complex, 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 that utilizes ACT, which I refer to as ACT-GMM.

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, the 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 an ACT change equation, which allows it to infer the EPA values of each entity in any particular social event given the others, and to calculate the likelihood of a social event when the EPA profiles of all entities are known. This equation is considered to be static and thus is not updated in any fashion during model inference, nor is it explicitly referenced in Figure B.1.

Language model component

All entities are assumed to be drawn from a simple language model. In the language model, actors and objects are 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. The variable Q is described in detail in the full article.

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. 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 values 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. 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.

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 easily 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 fundamental (pre-event) sentiment f. We could then provide a deterministic deflection score for the event by substituting these values into ACT's deflection equation. 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, as it accounts for possible transient meanings in the context of this particular social event beyond what we can account for with our mixture model. The distribution of deflection is assumed to be Laplacian, which makes model inference easier while still retaining the desired sociotheoretic meaning of deflection as a distance metric.

Bibliography

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford University Press, Oxford [etc.]. 3.1.1

Anderson, J. R., Matessa, M., and Lebiere, C. (1997). ACT-R: A theory of higher level cognition and its relation to visual attention. *Human-Computer Interaction*, 12(4):439–462. 2.1, 1, A.1.3

Arora, S., Ge, R., and Moitra, A. (2012). Learning Topic Models - Going beyond SVD. *arXiv:1204.1956*. A.2.2

Avenanti, A., Sirigu, A., and Aglioti, S. M. (2010). Racial bias reduces empathic sensorimotor resonance with other-race pain. *Current Biology*, 20(11):1018–1022. 7

Bamman, D. (2014). *People-Centric Natural Language Processing*. PhD thesis, Carnegie Mellon University. 2.2

Baym, N. K. (2010). *Personal Connections in the Digital Age*. Polity. 6

Bi, B., Tian, Y., Sismanis, Y., Balmin, A., and Cho, J. (2014). Scalable Topic-Specific Influence Analysis on Microblogs. A.2.2

Bishop, C. M. and others (2006). *Pattern recognition and machine learning*, volume 1. springer New York. 5.1.4

Blau, P. (1977). A macrosociological theory of social structure. *American journal of sociology*, pages 26–54. 6.2

Blau, P. M. (1974). Presidential Address: Parameters of Social Structure. *American Sociological Review*, 39(5):615–635. ArticleType: research-article / Full publication date: Oct., 1974 / Copyright 1974 American Sociological Association. 6.2

Blei, D. M., Griffiths, T. L., Jordan, M. I., and Tenenbaum, J. B. (2004). Hierarchical Topic Models and the Nested Chinese Restaurant Process. In *Advances in Neural Information Processing Systems*, page 2003. MIT Press. A.2.2

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. A.2.2

Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35. A.2.2

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022. 2.2, A.2.2

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM. A.2.1

Bonilla-Silva, E. (2006). *Racism without racists: Color-blind racism and the persistence of racial inequality in the United States*. Rowman & Littlefield Publishers. 7

Bourdieu, P. (1986). The forms of capital. *Handbook of theory and research for the sociology of education*, 241:258. 2.1, A.1.3

Bourdieu, P. (1990). *The logic of practice*. Stanford University Press. A.1.3

Boyd, D. and Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5):662–679. 8.2.3

Brewer, M. B. (1991). The social self: On being the same and different at the same time. *Personality and social psychology bulletin*, 17(5):475–482. 1

Bruns, A., Highfield, T., and Burgess, J. (2013). The arab spring and social media audiences english and arabic twitter users and their networks. *American Behavioral Scientist*, 57(7):871–898. 6

Calais Guerra, P. H., Veloso, A., Meira Jr, W., and Almeida, V. (2011). From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158. A.2.4

Cambria, E., Olsher, D., and Rajagopal, D. (2014). SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In *Twenty-eighth AAAI conference on artificial intelligence*. 2.2, A.2.4

Carley, K. (1990). *Content analysis*. The encyclopedia of language and linguistics. Edinburgh: Pergamon Press. 2.2, A.2.2

Carley, K. M. and Kaufer, D. S. (1993). Semantic connectivity: An approach for analyzing symbols in semantic networks. *Communication Theory*, 3(3):183–213. 2.2, A.2.2

Chang, J., Boyd-Graber, J. L., Gerrish, S., Wang, C., and Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. In *NIPS*, volume 22, pages 288–296. A.2.2

Chang, M.-W., Hsu, B.-J., Ma, H., Loynd, R., and Wang, K. (2014). E2e: An End-to-End Entity Linking System for Short and Noisy Text. *Making Sense of Microposts (# Microposts2014)*. A.2.1

Chapelle, O., Schlkopf, B., Zien, A., and others (2006). Semi-supervised learning. 4.1.3

Charniak, E. (1996). *Statistical language learning*. MIT press. B.2

Christensen, H. S. (2011). Political activities on the Internet: Slacktivism or political participation by other means? *First Monday*, 16(2). 7

Cikara, M. and Fiske, S. T. (2013). Their pain, our pleasure: stereotype content and schadenfreude. *Annals of the New York Academy of Sciences*, 1299(1):52–59. A.1.3

Cikara, M. and Van Bavel, J. J. (2014). The neuroscience of intergroup relations an integrative review. 9(3):245–274. 1, 7, A.1.1

Collins, A. M. and Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407. 2.1, A.1.3

Compton, R., Jurgens, D., and Allen, D. (2014). Geotagging one hundred million twitter accounts with total variation minimization. In *Big Data (Big Data), 2014 IEEE International*

Conference on, pages 393–401. IEEE. 5.1.2

Comunello, F. and Anzera, G. (2012). Will the revolution be tweeted? a conceptual framework for understanding the social media and the arab spring. *Islam and ChristianMuslim Relations*, 23(4):453–470. 6

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. 3, 7, A.1.1

de Oliveira, D. M., Laender, A. H., Veloso, A., and da Silva, A. S. (2013). FS-NER: A Lightweight Filter-stream Approach to Named Entity Recognition on Twitter Data. In *Proceedings of the 22Nd International Conference on World Wide Web Companion*, WWW '13 Companion, pages 597–604, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee. A.2.1

DiMaggio, P. (1997). Culture and cognition. *Annual review of sociology*, 23(1):263–287. A.1.3

Dixon, J., Levine, M., Reicher, S., and Durrheim, K. (2012). Beyond prejudice: Are negative evaluations the problem and is getting us to like one another more the solution? *Behavioral and Brain Sciences*, 35(06):411–425. 1

Dong, L., Wei, F., Liu, S., Zhou, M., and Xu, K. (2014). A Statistical Parsing Framework for Sentiment Classification. *arXiv preprint arXiv:1401.6330*. 2.2, A.2.4

Dou, W., Yu, L., Wang, X., Ma, Z., and Ribarsky, W. (2013). HierarchicalTopics: Visually Exploring Large Text Collections Using Topic Hierarchies. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2002–2011. 2.2, A.2.3

DAndrade, R. (2001). A Cognitivists View of the Units Debate in Cultural Anthropology. *Cross-Cultural Research*, 35(2):242–257. 4.2

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. Diffusion of lexical change in social media. 9(11):e113114. 5.1.4, 6

Fader, A., Soderland, S., and Etzioni, O. (2011). Identifying relations for open information extraction. In *Proceedings of the Conference of Empirical Methods in Natural Language Processing (EMNLP '11)*, Edinburgh, Scotland, UK. 2.2, A.2.2

Ferragina, P. and Scaiella, U. (2010). Fast and accurate annotation of short texts with Wikipedia pages. *arXiv:1006.3498 [cs]*. arXiv: 1006.3498. A.2.1

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. 7, 8.2.1, A.1.1

Flatow, D., Naaman, M., Xie, K. E., Volkovich, Y., and Kanza, Y. (2014). On the Accuracy of Hyper-local Geotagging of Social Media Content. *arXiv preprint arXiv:1409.1461*. 5.1.2

Gall, M., Renders, J.-M., and Karstens, E. (2013). Who broke the news?: an analysis on first reports of news events. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 855–862. International World Wide Web Conferences Steering Committee. 6

Gao, H., Mahmud, J., Chen, J., Nichols, J., and Zhou, M. (2014). Modeling User Attitude toward Controversial Topics in Online Social Media. In *the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM 2014)*. A.2.4

Gattani, A., Lamba, D. S., Garera, N., Tiwari, M., Chai, X., Das, S., Subramaniam, S., Rajaraman, A., Harinarayan, V., and Doan, A. (2013). Entity extraction, linking, classification, and tagging for social media: a wikipedia-based approach. *Proceedings of the VLDB Endowment*, 6(11):1126–1137. A.2.1

Gelvin, J. (2015). *The Arab uprisings: what everyone needs to know*. Oxford University Press, 2nd edition. 1, 6, 1

Gilbert, N. and Troitzsch, K. (2005). *Simulation for the social scientist*. Open university press. 3.2

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for Twitter: annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA. Association for Computational Linguistics. A.2.1

Godsill, S. J., Doucet, A., and West, M. (2004). Monte Carlo Smoothing for Nonlinear Time Series. *Journal of the American Statistical Association*, 99(465):156–168. 5.1.4

Goldberg, A. (2011). Mapping Shared Understandings Using Relational Class Analysis: The Case of the Cultural Omnivore Reexamined1. *American Journal of Sociology*, 116(5):1397–1436. 8.1.1, A.1.3

Goldstone, J. A. (2011). Cross-class coalitions and the making of the arab revolts of 2011. *Swiss Political Science Review*, 17(4):457–462. 6

Goldstone, J. A. (2013). Bringing regimes back inexplaining success and failure in the middle east revolts of 2011. *Available at SSRN 2283655*. 6

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, page mps028. 3, A.2.4

Guerra, P. C., Meira Jr, W., and Cardie, C. (2014). Sentiment analysis on evolving social streams: How self-report imbalances can help. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 443–452. ACM. A.2.4

Guerra, P. H. C., Meira Jr, W., Cardie, C., and Kleinberg, R. (2013). A Measure of Polarization on Social Media Networks Based on Community Boundaries. In *Seventh International AAAI Conference on Weblogs and Social Media*. A.2.4

Guo, S., Chang, M.-W., and Kcman, E. (2013). To link or not to link? a study on end-to-end tweet entity linking. In *Proceedings of NAACL-HLT*, pages 1020–1030. A.2.1

Habib, M. B. and Keulen, M. (2013). A generic open world named entity disambiguation approach for tweets. 2.2

Hajishirzi, H., Zilles, L., Weld, D. S., and Zettlemoyer, L. S. (2013). Joint Coreference Resolution and Named-Entity Linking with Multi-Pass Sieves. In *EMNLP*, pages 289–299. 4.2

Hall, E. V., Phillips, K. W., and Townsend, S. S. (2015). A rose by any other name?: The consequences of subtyping African-Americans from Blacks. *Journal of Experimental Social Psychology*, 56:183–190. 7

Han, J., Wang, C., and El-Kishky, A. (2014). Bringing Structure to Text: Mining Phrases, Entities, Topics, and Hierarchies. In *Proceedings of the 20th ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining, KDD '14, pages 1968–1968, New York, NY, USA. ACM. A.2.1, A.2.2

Harris, F. C. and Lieberman, R. C. (2015). Racial Inequality after Racism: How Institutions Hold Back African Americans. *Foreign Affairs*, 94:9. 7

Heise, D. R. (2001). Project magellan: Collecting cross-cultural affective meanings via the internet. *Electronic Journal of Sociology*, 5(3). 4.1.2

Heise, D. R. (2007). *Expressive Order*. Springer. 1, 1, 2.1, 5.1.3, 8, A.1.1, A.1.2, A.1.3

Heise, D. R. (2010a). INTERACT: Introduction and Software. 2

Heise, D. R. (2010b). *Surveying cultures: Discovering shared conceptions and sentiments*. John Wiley & Sons. 3, A.1.1

Heise, D. R. and MacKinnon, N. J. (2010). *Self, identity, and social institutions*. Palgrave Macmillan. 1, 2.1, 3.1.1, 4, 1, 4.1.4, 4.2, 4.3, 8.1.2, 8.2.1, A.1.1, A.1.3

Hewstone, M., Rubin, M., and Willis, H. (2002). Intergroup bias. *Annual review of psychology*, 53(1):575–604. 1, 2.1, A.1.1, 1

Hoang, T.-A., Cohen, W. W., and Lim, E.-P. (2014). On modeling community behaviors and sentiments in microblogging. SIAM. 2.2, A.2.4

Hoey, J. and Schröder, T. Bayesian affect control theory of self. In *Proceedings of the AAAI Conference on Artificial Intelligence*. A.1.3

Hoey, J., Schroder, T., and Alhothali, A. (2013a). Bayesian Affect Control Theory. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 166–172. 3.1.1, A.1.3

Hoey, J., Schröeder, T., and Alhothali, A. (2013b). Affect Control Processes: Intelligent Affective Interaction using a Partially Observable Markov Decision Process. *arXiv:1306.5279 [cs]*. arXiv: 1306.5279. A.1.3

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098. 3.1.1, 5.1.4, 5.2

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. A.2.1

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. A.2.1

Hong, L. and Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88. ACM. A.2.2

Hu, Z., Wang, C., Yao, J., Xing, E., Yin, H., and Cui, B. (2013). Community Specific Temporal Topic Discovery from Social Media. *arXiv preprint arXiv:1312.0860*. A.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*. A.2.4

Joseph, K., Carley, K. M., Filonuk, D., Morgan, G. P., and Pfeffer, J. (2014a). Arab spring: from

newspaper data to forecasting. *Social Network Analysis and Mining*, 4(1):1–17. 6, 6.1.1

Joseph, K., Carley, K. M., and Hong, J. I. (2014b). Check-ins in "Blau Space": Applying Blau's Macrosociological Theory to Foursquare Check-ins from New York City. *ACM Trans. Intell. Syst. Technol.*, 5(3):46:1–46:22. A.2.2

Joseph, K., Morgan, G. P., Martin, M. K., and Carley, K. M. (2014c). On the Coevolution of Stereotype, Culture, and Social Relationships An Agent-Based Model. *Social Science Computer Review*, 32(3):295–311. 3.2

Joseph, K., Wei, W., Benigni, M., and Carley, K. M. (sub.). Inferring affective meaning from text using Affect Control Theory and a probabilistic graphical model. *Journal of Mathematical Sociology*. 3.1.1, 3.2

Kim, D. I. and Sudderth, E. B. (2011). The doubly correlated nonparametric topic model. In *Advances in Neural Information Processing Systems*, pages 1980–1988. A.2.2

Kim, S., Li, F., Lebanon, G., and Essa, I. (2012). Beyond sentiment: The manifold of human emotions. *arXiv preprint arXiv:1202.1568*. 2.2, A.2.4

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press. 5.1.3

Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., and Smith, N. A. (2014). A dependency parser for tweets. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, to appear*. 4.1.2, A.2.1

Krishnan, V. and Eisenstein, J. (2014). Unsupervised Induction of Signed Social Networks from Content and Structure. *arXiv preprint arXiv:1411.4351*. 2.2, A.2.4

Kroeber, A. L. and Kluckhohn, C. (1952). Culture: A critical review of concepts and definitions. *Papers. Peabody Museum of Archaeology & Ethnology, Harvard University*. A.1.3

Kbler, S., McDonald, R., and Nivre, J. (2009). Dependency parsing. *Synthesis Lectures on Human Language Technologies*, 1(1):1–127. 2.2, 4.1.2, 5.1.2, A.2.2, B.1

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 4.1.3

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., and Jurafsky, D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916. 2.2, 4.1.2, A.2.1

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. A.2.3

Leetaru, K. (2011). Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, 16(9). B.1

Li, H., Chen, Y., Ji, H., Muresan, S., and Zheng, D. (2012a). Combining social cognitive theories with linguistic features for multi-genre sentiment analysis. In *In Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation*. A.2.4

Li, W., Blei, D., and McCallum, A. (2012b). Nonparametric Bayes Pachinko Allocation. *arXiv:1206.5270*. A.2.2

Lin, T., Etzioni, O., and others (2012). No noun phrase left behind: detecting and typing unlink-

able 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. A.2.1

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*. A.2.4

Liu, Y., Niculescu-Mizil, A., and Gryc, W. (2009). Topic-link LDA: joint models of topic and author community. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 665–672. ACM. A.2.2

Lizardo, O. (2006). How Cultural Tastes Shape Personal Networks. *American Sociological Review*, 71(5):778–807. 3, A.1.3

Lizardo, O. (2011). Cultural correlates of ego-network closure. *Sociological Perspectives*, 54(3):479–487. 3, A.1.3

Lizardo, O. (2014). Omnivorousness as the bridging of cultural holes: A measurement strategy. *Theory and Society*, 43(3-4):395–419. 3, A.1.3

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. 2.1, 7, A.1.3

Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I., and Boyd, D. (2011). The revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. *International Journal of Communication*, 5:1375–1405. 6

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 *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 142–150. Association for Computational Linguistics. A.2.4

Mahmud, J., Nichols, J., and Drews, C. (2014). Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):47. 5.1.2

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. 5

McPherson, M. (2004). A Blau space primer: prolegomenon to an ecology of affiliation. *Industrial and Corporate Change*, 13(1):263–280. 6.2

Mead, G. H. (1925). The Genesis of the Self and Social Control. *International Journal of Ethics*, 35(3):251–277. ArticleType: research-article / Full publication date: Apr., 1925 / Copyright 1925 The University of Chicago Press. A.1.3

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. 2.2, A.2.1

Meij, E., Weerkamp, W., and de Rijke, M. (2012). Adding semantics to microblog posts. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 563–572. ACM. A.2.1

Mikolov, T., Yih, W.-t., and Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. In *HLT-NAACL*, pages 746–751. Citeseer. A.2.3

Miles, A. (2014). Addressing the Problem of Cultural Anchoring An Identity-Based Model of Culture in Action. *Social Psychology Quarterly*, 77(2):210–227. 1, 2.1, 2, 8.2.1, A.1.3

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41. 4

Mimno, D. and Blei, D. (2011). Bayesian checking for topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 227–237, Stroudsburg, PA, USA. Association for Computational Linguistics. A.2.2

Mimno, D. and McCallum, A. (2012). Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression. *arXiv:1206.3278*. A.2.2

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics. 4.1.3

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., and Rosenquist, J. N. (2011). Understanding the Demographics of Twitter Users. *ICWSM*, 11:5th. 5.1.2

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. (2015a). Never-ending learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*. 2.2, A.2.1, A.2.2

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. (2015b). Never-Ending Learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*. 5.1.2

Mohr, J. W. and Bogdanov, P. (2013). IntroductionTopic models: What they are and why they matter. *Poetics*, 41(6):545–569. A.2.2

Moro, A., Raganato, A., and Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation: A Unified Approach. *Transactions of the Association for Computational Linguistics*, 2. 2.2, 8.2.2

Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. M. (2013). Is the sample good enough? comparing data from twitters streaming api with twitters firehose. *Proceedings of ICWSM*. 8.2.3

Mukherjee, S. (2014). Joint Author Sentiment Topic Model. In *SDM*. 2.2, A.2.2, A.2.4

Mukherjee, S. and Bhattacharyya, P. (2013). Sentiment Analysis : A Literature Survey. *arXiv:1304.4520 [cs]*. arXiv: 1304.4520. A.2.4

Neiswanger, W., Wang, C., and Xing, E. (2013). Asymptotically exact, embarrassingly parallel MCMC. *arXiv preprint arXiv:1311.4780*. A.2.2

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. A.2.2

Osgood, C. E. (1969). On the whys and wherefores of E, P, and A. *Journal of Personality and*

Social Psychology, 12(3):194. 2.1, A.1.1, B.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*. 4.1.2, A.2.1

OConnor, B., Bamman, D., and Smith, N. A. (2011). Computational Text Analysis for Social Science: Model Assumptions and Complexity. *public health*, 41(42):43. 3

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. 1, 8

Paluck, E. L. and Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research and practice. *Annual review of psychology*, 60:339–367. 1

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135. 2.2, A.2.4, B.1

Pfeffer, J. and Carley, K. M. (2012). Rapid modeling and analyzing networks extracted from prestructured news articles. *Computational and Mathematical Organization Theory*, 18(3):280–299. 6

Rabinovich, M. and Blei, D. M. (2014). The Inverse Regression Topic Model. In *ICML '14*. A.2.2

Ratnaparkhi, A. and others (1996). A maximum entropy model for part-of-speech tagging. In *Proceedings of the conference on empirical methods in natural language processing*, volume 1, pages 133–142. Philadelphia, PA. 2.2

Ridgeway, C. L. and Kricheli-Katz, T. (2013). Intersecting Cultural Beliefs in Social Relations Gender, Race, and Class Binds and Freedoms. *Gender & Society*, 27(3):294–318. 3.1.1

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. 2.2, A.2.1, A.2.2

Roberts, M. E., Stewart, B. M., and Airoldi, E. M. (2013). Structural Topic Models. Technical report, Working paper. A.2.2

Robinson, D. T., Smith-Lovin, L., and Wisecup, A. K. (2006). *Affect control theory*. Springer. 1, 2.1, A.1.1

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, 2.1, 8.2.1, A.1.1

Rumelhart, D. (1978). *Schemata: The building blocks of cognition*. Center for Human Information Processing, University of California, San Diego. 2.1, A.1.3

Ruths, D. and Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, 346(6213):1063–1064. 8.2.3

Sachan, M., Dubey, A., Srivastava, S., Xing, E. P., and Hovy, E. (2014). Spatial Compactness Meets Topical Consistency: Jointly Modeling Links and Content for Community Detection. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 503–512, New York, NY, USA. ACM. A.2.2

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. 7, 8

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, volume 12, pages 44–49. Citeseer. 4.1.2

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. 2

Simon, H. A. (1987). Bounded rationality. *The New Palgrave: utility and probability*, pages 15–18. A.1.3

Singh, S., Subramanya, A., Pereira, F., and McCallum, A. (2012). Wikilinks: A Large-scale Cross-Document Coreference Corpus Labeled via Links to Wikipedia. Technical Report UM-CS-2012-015. 4.1.2

Smith-Lovin, L. (1987). The affective control of events within settings. *Journal of Mathematical Sociology*, 13(1-2):71–101. 7, A.1.1

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. 3.1.2, A.1.1

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. 2.2, 4.1.2, A.2.1

Srivastava, S. B. and Banaji, M. R. (2011). Culture, Cognition, and Collaborative Networks in Organizations. *American Sociological Review*, 76(2):207–233. 7, A.1.3

Starbird, K. and Palen, L. (2012). (how) will the revolution be retweeted?: information diffusion and the 2011 egyptian uprising. In *Proceedings of the acm 2012 conference on computer supported cooperative work*, pages 7–16. ACM. 6

Stryker, S. From mead to a structural symbolic interactionism and beyond. 34:15–31. A.1.3

Stryker, S. (1980). *Symbolic interactionism: A social structural version*. Benjamin-Cummings Publishing Company. 5.1.3

Swidler, A. (1986). Culture in action: Symbols and strategies. *American sociological review*, pages 273–286. 2.1, A.1.3

Taddy, M. (2013). Measuring Political Sentiment on Twitter: Factor Optimal Design for Multinomial Inverse Regression. *Technometrics*, 55(4):415–425. A.2.4

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. A.1.1

Takahashi, Y., Utsuro, T., Yoshioka, M., Kando, N., Fukuhara, T., Nakagawa, H., and Kiyota, Y. (2012). Applying a Burst Model to Detect Bursty Topics in a Topic Model. In Isahara, H. and Kanzaki, K., editors, *Advances in Natural Language Processing*, volume 7614 of *Lecture Notes in Computer Science*, pages 239–249. Springer Berlin / Heidelberg. A.2.2

Taskar, B., Guestrin, C., and Koller, D. (2003). Max-Margin Markov Networks. In *Advances in Neural Information Processing Systems*, page None. 4.1.3

Tufekci, Z. (2014). Big Questions for Social Media Big Data: Representativeness, Validity and

Other Methodological Pitfalls. In *ICWSM 14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media.* 8.2.3

Tufekci, Z. and Wilson, C. (2012). Social media and the decision to participate in political protest: Observations from tahrir square. *Journal of Communication*, 62(2):363–379. 6

Vaisey, S. (2009). Motivation and Justification: A DualProcess Model of Culture in Action. *American Journal of Sociology*, 114(6):1675–1715. A.1.3

Van Bavel, J. J., Packer, D. J., and Cunningham, W. A. (2008). The Neural Substrates of In-Group Bias A Functional Magnetic Resonance Imaging Investigation. *Psychological Science*, 19(11):1131–1139. 7

Vrandei, D. and Krtzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85. A.2.1

Wagner, C., Asur, S., and Hailpern, J. (2013). Religious politicians and creative photographers: Automatic user categorization in twitter. In *Social Computing (SocialCom), 2013 International Conference on*, pages 303–310. IEEE. 5.1.2

Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1105–1112, New York, NY, USA. ACM. A.2.2

Wang, C., Chakrabarti, K., Cheng, T., and Chaudhuri, S. (2012). Targeted disambiguation of ad-hoc, homogeneous sets of named entities. In *Proceedings of the 21st international conference on World Wide Web*, pages 719–728. ACM. 2.2, A.2.1

Wang, S. and Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 90–94. Association for Computational Linguistics. A.2.4

Williams, P. J. (1997). *Seeing a Color-Blind Future: The Paradox of Race*. Macmillan. 7

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. A.2.1, A.2.2

Yin, Z., Cao, L., Han, J., Zhai, C., and Huang, T. (2011). LPTA: A Probabilistic Model for Latent Periodic Topic Analysis. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 904 –913. A.2.2

Zhu, J. and Xing, E. P. (2009). Maximum Entropy Discrimination Markov Networks. *J. Mach. Learn. Res.*, 10:2531–2569. 4.1.3

Zhu, L., Galstyan, A., Cheng, J., and Lerman, K. (2014). Tripartite Graph Clustering for Dynamic Sentiment Analysis on Social Media. *arXiv preprint arXiv:1402.6010*. A.2.4