Girls rule, boys drool: Extracting semantic and affective stereotypes on Twitter

Kenneth Joseph
Northeastern University*
Boston, MA
k.joseph@northeastern.edu

Wei Wei
Carnegie Mellon University
Pittsburgh, PA
weiwei@cs.cmu.edu

Kathleen M. Carley
Carnegie Mellon University
Pittsburgh, PA
kathleen.carley@cs.cmu.edu

ABSTRACT

Social identities carry widely agreed upon meanings, called stereotypes, that have important effects on social processes. We develop a method to extract the stereotypes of a particular population of Twitter users. Our model is grounded in social theory on stereotypes as both identities' affective meanings and their semantic relationships to each other. We apply our model to a dataset of 45K Twitter users who actively tweeted about the Michael Brown and Eric Garner tragedies. This case study furthers our understanding of both the stereotypes present for those who actively discussed these tragedies online as well as the structure of stereotypes in wider populations both online and off.

Keywords

computational social science; social psychology; identity; stereotype; Twitter

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences

1. INTRODUCTION

A social identity is a word or phrase that defines a type, group or class of individuals [27]. Our social identities have a profound impact on our lives. For example, those who choose (or are given) the identity of a woman or an African American have significantly reduced employment opportunities as compared to their white, male counterparts [5]. The way in which social identities impact our lives on an interpersonal level[1] can be broken into at least two major phenomena. The first is the process by which we select identities for ourselves and others. The second is the process by which we, given these "identity labelings", determine how to behave towards those others.

It is widely believed that both how we select identities and how these selections impact our behavior are determined in large part by the *stereotypes*, or meanings, attached to each social identity. Much less agreed upon is exactly how stereotypes impact our labeling of others and the behaviors we choose to enact. *Affect Control Theory (ACT)* [25] is one of the few theories that provides a predictive model for how stereotypes are used to label others and how these labelings impact interpersonal behavior. Over the past several decades, the theory has seen significant use by sociologists, social psychologists [25, 26, 29, 56, 50], and cognitive psychologists [55, 54] interested in interpersonal interaction. It has also caught the attention of computer scientists, who have applied it to allow robots to perform more human-like actions in interpersonal settings [31, 30] and to the study of social interactions described in text [1, 34].

In order to predict labeling and social behavior, ACT relies on the assumption that stereotypes can be quantified *affectively*. In other words, ACT assumes that stereotypes are defined by how people *feel* about an identity [50]. A core component of this assumption, supported by empirical evidence [27], is that within a particular national culture, these affective stereotypes are stable across time. For example, the theory assumes it is very likely that almost all Americans, from those living in the 1920s to those living today, agree that villains are bad and that heroes are good. This consistency has allowed Affect Control scholars to collect affective stereotypes for hundreds of identities representative of entire national cultures using rigorous, time-intensive survey techniques on relatively small sample sizes [26, 28].

There are, however, two well-known shortcomings of ACT. First, while many affective stereotypes are stable and culturally consistent, it has also been shown that in sub-populations where particular identities are highly relevant to the population's own identity (e.g. the identity "professor" at an academic conference), there may be systematic differences in affective stereotypes for these identities within the sub-population as compared to the nation at large [56, 59]. These variations may lead ACT's predictive models to struggle when applied to a particular population of interest. Given the difficulty and expenses associated with surveys used to collect affective stereotypes, they can also be hard to correct.

Second, affective stereotypes alone are insufficient for the prediction of both identity labeling and behavior in interpersonal settings [25, 28, 34]. This is because in addition to affective stereotypes, *semantic relationships* between identities, referred to here as *semantic stereotypes*, also exist. For example, regardless of affect, when we think of teachers, we generally think of them as interacting with students. These semantic stereotypes have received little empirical attention from the ACT community. ACT researchers instead largely assume that identities can be grouped into *institutional structures* - that is, into clusters of identities that align with differ-

*The majority of this work was completed while the author was a graduate student at Carnegie Mellon University

[1]That is, not on a systemic level, where a whole new space of phenomena arises- see, e.g. [10]

CSCW '17, February 25-March 01, 2017, Portland, OR, USA

ISBN 978-1-4503-4335-0/17/03…$15.00

DOI: http://dx.doi.org/10.1145/2998181.2998187

ent institutionalized settings, like home, school and work, that are encountered in everyday life [29, 35]. Like affective stereotypes, these institutional structures are assumed to be relatively static and consistent within a particular national culture, even though little has been done to test this assertion. Further, little has been done to understand how identities interrelate based on the way individuals move across settings aligned with different institutions.

The present work takes aim at these two problems by developing a method to rapidly infer the affective and semantic stereotypes of a particular population using Twitter data. To infer affective stereotypes, we develop a slight generalization of current approaches to extracting affective stereotypes from text [34, 1], the first to of its kind to apply to Twitter data. In doing so, we create a small but important connection between NLP methods related to ACT and the broader literature on sentiment analysis on Twitter. To infer semantic stereotypes, we rely on a blend of existing work from cognitive psychologists [39, 23] and NLP scholars [7] that emphasizes the *network structure* of semantic stereotypes, rather than the clustering of identities into institutions.

While the method we develop thus relies heavily on prior work, we show how it can be leveraged to provide important insights into the affective and semantic stereotypes of a population. We apply our approach to tweets from a set of 45K Twitter users actively engaged in discussions about the Eric Garner and Michael Brown tragedies between 2013 and 2015 (hereafter, the *EG/MB population*). Our results provide a novel view into the stereotypes of this population, and to the extent we show our findings generalize, to how stereotypes may be structured in broader populations as well.

2. LITERATURE REVIEW

A large and growing literature exists on how stereotypes impact language - for a review, we point the reader to [6]. Recent research has also focused on developing novel NLP techniques to study stereotyping and its impacts [4, 22, 1, 34, 9, 11]. These articles suggest there is much to be gained from utilizing language and NLP methods to study stereotyping.

The primary goal of the present work is to leverage largely existing NLP methods to improve the measurements and theoretical underpinnings of Affect Control Theory. Our efforts thus complements prior work, showing how stereotypes derived from language can be utilized in a social theory which explains how stereotypes actually impact social behavior. In the sections below, we review work related to how we extract semantic and affective stereotypes, touching on underlying concepts in social and cognitive psychology as well as related NLP literature.

2.1 Semantic Relations as Stereotypes

Heise and MacKinnon [29] were the first to provide empirical evidence of institutional structure in the semantic relationships between identities. The authors used semantic network analysis of dictionaries and clustered these networks into institutions. More recently, Joseph et al. [35] studied semantic clustering of identities on Twitter using latent Dirichlet allocation [35]. They also found evidence of stable institutions in addition to hints of more dynamic forces shaping the "topics" of identities they uncovered.

The present work extends these efforts in two ways. First, prior work has made no attempt to compare multiple populations to understand how consistent semantic clusterings centered around institutions might in fact be. In the present work, we compare results between two populations to get a better sense of this level of stability. Second, prior work fails to focus on semantic associations between identities at the pairwise level. Consequently, little has been done to understand how identities may fall into important positions at the boundaries between institutions. In the present work, we leverage a representation of semantic stereotypes as a network of cognitive associations between identities in order to address this shortcoming.

Modeling semantic stereotypes as a network provides two additional advantages over modeling identities as belonging to specific institutions. First, such networks can always be clustered to reveal any institutional structure that might exist. Second, and more importantly, the idea of semantic stereotypes as a network of cognitive associations is better aligned with existing observations of how the human mind functions [16, 3, 2]. Indeed, cognitive psychologists have for some time modeled stereotypes in explicitly such a network fashion [39, 23] and have leveraged such models in research on interpersonal settings.

Parallel Constraint Satisfaction Models (PCSMs), have increasingly become the model of choice for researchers leveraging semantic associations in this way to study social phenomena. In PCSMs, nodes represent identities and links represent both positive and negative cognitive associations between them. Cognitive activation is assumed to spread along positive links and to be inhibited by negative links in a way that leads individuals to label people in certain ways or engage in certain types of behaviors. PCSMs have shown promise in simulation studies in explaining certain poorly understood social psychological concepts including intersectionality [23] and social priming [55].

No method currently exists to parameterize PCSMs beyond handwiring of links by researchers, making them difficult to use beyond small-scale simulation studies. Conveniently, however, NLP scholars have long studied techniques to extract networks of semantic associations from text [18, 43, 12], and one can draw a direct parallel from the assumptions of PCSMs to the NLP literature. Specifically, the Correlated Topic Model (CTM) [7] presents a method to extract correlations between topics in the form of a Gaussian covariance matrix. If one assumes a Gaussian distribution over cognitive activation at each node in a PCSM, PCSMs can be thought of as Gaussian Markov Random Fields, which in turn can be represented via the inverse of a Gaussian covariance matrix.

The problem of extracting the semantic network of a PCSM from text thus reduces to removing the assumption of topics from the CTM and applying it to Twitter data, assuming Twitter users are "bags of identities". From here, we can obtain a covariance matrix which can be transformed into a PCSM at will, representing which pairs of identities tend to be used by similar sets of individuals. This is the approach taken in the present work.

2.2 Affective Meanings as Stereotypes

At its most general level, ACT is a social psychological model of how humans interpret and behave within interpersonal social settings.[2] There are three assumptions of ACT that are relevant to the present work. First, ACT assumes a particular measurement system for affective stereotypes of identities, the behaviors these identities engage in, and modifiers (e.g. "bad") that can be used to describe identities. Affective meanings of these entities are defined in a three dimensional *EPA* space with axes entitled *E*valuation (goodness/badness of an identity), *P*otency (strength/weakness of an identity) and *A*ctivity (activeness/passiveness of an identity), each spanning the range of -4.3 to +4.3. The position of an entity within this space is called its *EPA profile*.

Second, ACT assumes that humans make decisions in interpersonal settings based on the concept of *deflection*. Deflection expresses how "expected" a particular labeling of an individual is or

[2]For a high-level overview, we suggest [50], for a more technical perspective, [31].

how expected a particular action engaged in by that individual is. Finally, ACT assumes that deflection is caused by the *social events* we observe or engage in during interpersonal interaction. A social event is an interaction in which an *actor* identity enacts a *behavior* on an *object* identity [27].

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

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

ACT provides a regression equation that is used to determine the elements of the post-event transient f' as a function of f. Mathematically, $f' = \mathcal{M}g(f)$, where $g(f)$ is a kx1 vector of covariates (e.g. $\begin{bmatrix} 1 & a_e & \dots & b_e o_e & \dots & a_e b_e o_e \end{bmatrix}$) and $\mathcal{M}$ is a 9xk matrix specifying 9 different sets of regression coefficients, one for each element of f'. Importantly, $g(f)$ consists of only linear combinations of the elements of f. The actual covariates $g(f)$ and coefficients ($\mathcal{M}$) used in this model are estimated via survey data; we refer the reader to [45] for details.

Assuming the form of $g(f)$ and the values in $\mathcal{M}$ are given, as we do here, the post-event transient can be constructed as follows, where $\mathcal{M}_x$ represents row x of the coefficient matrix:

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

ACT also allows for a regression model in which modifiers change how social events impact perception by changing the meaning of an the actor or object (e.g. a "bad teacher" is different than a "teacher") - for details, we refer the reader to [27].

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

$$deflection = \sum_j^9 (f_j - f'_j)^2 = \sum_j^9 (f_j - \mathcal{M}_j \cdot g(f))^2 \qquad (1)$$

A high deflection score for an event means the affective meanings of the actor, behavior and/or object have changed dramatically. This signifies that this event was relatively unlikely to have occurred given pre-existing affective stereotypes of the entities within the event. ACT can be used as a predictive model of identity labeling or behavior selection because it can be used to determine the optimal (in the sense that the event becomes "most expected", or has the lowest deflection) EPA profile of the actor, behavior or object given information on the other two. One can then look up the identity or behavior closest to this EPA profile to complete the prediction. In a related fashion, social events that are discussed by an individual on Twitter can be used to infer the individual's affective stereotypes by assuming the individual tends to express statements about identities that have low deflection.

A crucial observation in the present work is that there is a straightforward way to include additional factors into the deflection model beyond social events. While previous work using ACT for text data has relied solely on this existing social event framework to extract EPA profiles from text [1, 35], the sentiment analysis literature contains a vast array of additional constraining factors on sentimental meaning of words in text. In the present work, we introduce a framework for including these additional factors into the deflection model.

Sentiment mining of Twitter is a particularly popular area of research [51, 32, 40, 53]. Recently, scholars have focused on assessing the affective meaning of a concept or expression across an entire corpus of Twitter data [36, 58, 61], a task called *concept-level sentiment mining*. Notably, Chen et al. [15] use a graph-based algorithm to extract affective meanings of n-grams within text. Their model utilizes the concepts of consistent vs. inconsistent affective relationships, a conceptual model we also find appealing and make use of in the present work.

Similar approaches have considered the problem of target dependent sentiment analysis for Twitter data [33], which focuses on whether or not a particular tweet (as opposed to the entire corpus) is positive or negatively focused on a particular concept. Our "affective constraint model", to be described below, is related to several of these efforts in our use of a dependency parse to extract additional sentiment information from the text [20, 65], an approach also utilized in certain concept-level methods as well [49].

In sum, a plethora of methods exist to extract sentiment information from Twitter data. These methods range from simple word-counting (e.g. with LIWC [47]) to more complex, neural models (e.g. [58]). In the present work, our goal was to develop a simple model that was aligned with the core theoretical tenets of ACT (a three dimensional sentiment profile, social events and deflection) that could also incorporate additional sentiment information held within tweets. Our approach to doing so was to generalize existing NLP models incorporating ACT [1, 35] to also utilize keyword, emoji and emoticon-based sentiment expressions that are particularly effective for sentiment mining on Twitter. We thus take a small but important step towards unifying applied sentiment analysis with a major theoretical model of affective stereotypes.

3. DATA

In the present work, we focus on stereotypes of a manually constructed list of 310 identities of interest. These identities were selected by us based on their frequency of use within our dataset as well as their importance to prior research on identity [17, 35]. Some of these identities are of particular interest to the EG/MB population (e.g. "police"), while others represent more generic identities which we felt were important in understanding structures of semantic stereotypes within our data.

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

In October 2015, we re-collected data for all 250K users who sent at least one geotagged tweet.[3] The present work focuses on a subset of these users for which we were able to obtain their full set of tweets from 2013-2015, who sent between 250 and 10K tweets, had less than 50K followers, and who had at least 50 tweets that contained one or more of the 310 identities of interest. We did not consider retweets or tweets with fewer than five unique non-punctuation terms.

In addition to the Twitter dataset and identity list described, we also leverage two large dictionaries of words matched to their EPA profiles, collected via survey methods. The first is a list of approximately 2K EPA profiles for identities, emotions, behaviors and modifiers collected by ACT scholars [57].[4] The second is

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

[4]While a description of the particular dataset we use is not currently available, the reader can visit http://www.indiana.edu/~socpsy/ACT for a variety of datasets collected in the same fashion.

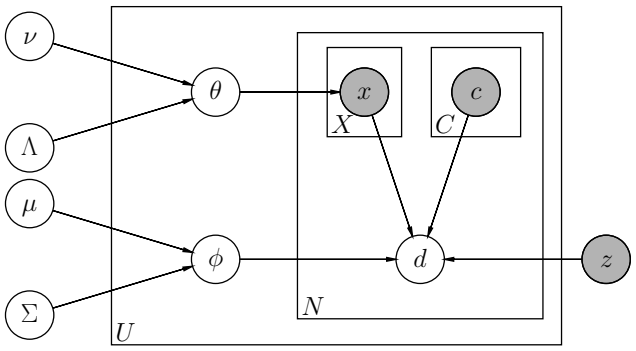

Figure 1: A graphical model of our method.

a list of approximately 14K EPA profiles for a variety of words collected via Amazon Mechanical Turk [62]. We also leverage EMOLEX [44], a survey dataset linking words expressed on Twitter to emotions. The emotions words are transformed into EPA values by mapping these emotion words into our two EPA dictionaries. Finally, we utilize the Emoji dataset from [38] to leverage sentiment expressed via these tokens. With the exception of the raw data from [57], all of these datasets are publicly available; all code and data from the present work can be found at https://github.com/kennyjoseph/twitter_stereotype_extraction.

4. MODEL

The model we use to infer semantic and affective stereotypes is presented in Figure 1 (without hyperparameters). For each user u in our dataset U, we have N_u tweets. Each tweet n for user u contains a (possibly empty) set of identities of interest found in the tweet's text, $X_{u,n}$. We consider any noun or adjective whose surface form is in our set of identities of interest to be an identity. If none are found, we ignore the tweet.

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

The model has two components. The first, consisting of parameters ν, Λ, and θ, is used to infer semantic relationships between identities. The parameter matrix θ estimates the extent to which a user tweets about each identity, θ_u is a row of this matrix defining values for the user u. Following the language of PCSMs (and [21]), we refer to values in θ as *activation scores*. The parameters ν and Λ define mean and covariance parameters over these activation scores, respectively. Parameterization follows the CTM, with a logistic normal prior over θ. We perform a fully Bayesian analysis, putting a conjugate Normal Inverse-Wishart (NIW) prior over Λ and ν. Formally, this portion of the model can be defined as follows:

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

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

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

Here, μ provides the mean affective ratings for each sentiment dimension of each identity and Σ provides the associated covariance matrix. The parameter matrix ϕ gives per-user values for each sentiment dimension for each user; ϕ_{u,i_e} represents the element of ϕ corresponding to the $e$valuative dimension of the ith identity for user u. The core of the sentiment model is the development of the probability distribution of d.[5] This variable represents the $d$eflection of a particular tweet. Subscripts are given above for q, X, ϕ and C to emphasize that the parameterization of d is unique for each tweet and independent across users.

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

4.1 Sentiment Constraint Extraction

We construct $q_{u,n}$ by extracting four types of semantic constraints - "clause-level", "emoji", "social event", and "social action" constraints. In order for us to do so, two preprocessing steps are necessary. First, each tweet is dependency parsed using a Twitter-specific dependency parser [37]. Second, each tweet is run through a classifier [35] to determine if any elements of C are also identities (that are not in our set of interest).[6] We can then proceed to extract sentiment constraints, and we describe each constraint below. Note that constraints are formulated as a quadratic function of elements of ϕ_u, this is important for efficient inference. The u subscript is made implicit in this section in order to ease notation. Similarly, we will use X to represent $X_{u,n}$ and C for $C_{u,n}$.

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

If an emoji in our dictionary is found in a tweet, a constraint is added to the evaluative dimension of each identity in X with the

[5]Note that while Equation 1 is deterministic, we follow prior work [34, 31] in the assumption that it is more appropriate to think of deflection as a random variable, influenced by other unknown factors.
[6]We do not use this classifier to extract identities of interest in order to mitigate any biases it may induce.

affective value for that emoji. For each element of X, a constraint of the form $(\phi_{x_e} - z_{j_e})^2$, where j is the emoji's index in z, is added.

We then extract social event constraints. Our approach follows typical extraction of Subject, Verb, Object triplets using a dependency parse - we look for verbs that are in C that have a direct subject and object which are both identities, at least one of which is in X and both of which are in X or C. We also extract modifier terms, elements of C which are adjectives and direct descendants in the dependency parse of these identities, and apply the ACT modifier equation in these cases.

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

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

Given the form of this pre-event transient, the full social event constraint can then be specified as in Equation 1.

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

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

4.2 Inference

Model inference is performed via Gibbs Sampling. The Gibbs sampler can be split into two parallelizable components, one which infers parameters for the semantic portion of the model and one which infers parameters for the affective portion of model. The Gibbs sampler for the semantic portion is derived directly from the work of Chen et al. [14], who develop an auxiliary variable method to perform Gibbs sampling on the CTM. Beyond the removal of topics from their model, the inference procedure is identical and is thus not covered here.

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

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

$$p(\phi_{i_e}|\mu, \Sigma, \phi_{\neg i_e}) \sim$$
$$\mathcal{N}\left(\mu_{i_e} - \Sigma^{-1}_{i_e,i_e}\Sigma^{-1}_{i_e,\neg i_e}(\phi_{\neg i_e} - \mu_{\neg i_e}), \Sigma^{-1}_{i_e,i_e}\right) \quad (2)$$

The likelihood is also Gaussian, though the derivation requires slightly more thought. Recall that $p(d)$ is Laplace distributed. Let us assume, as we will throughout, that $\beta = 1$. As noted above, $q_{u,n}(\phi_u, X_{u,n}, C_{u,n}, z)$ deterministically returns an equation that represents quadratic constraints on elements of ϕ. Given fixed values (for a given Gibbs step) for all parameters except ϕ_{i_e}, and noting we are only interested in a sampling distribution for this particular element, we can ignore the value of d_n and say, for a particular tweet from a particular user, that $p(d_n|\phi, .) \propto \exp(\frac{-|q_n(\phi, X_n, C_n, z)|}{2})$.

Recalling that $q_n(\phi, X_n, C_n, z)$ is quadratic in ϕ_{i_e} by construction, it should be clear that with some rearranging of variables and by addressing the absolute value, we are left with a Gaussian distribution in ϕ_{i_e}. A formal proof of this is given in early work on NLP models of ACT [34], we skip the proof here in the interest of space. The conditional sampling distribution of ϕ_{i_e} thus amounts to a conjugate normal update, with the only distinction being that each $p(d_n)$ has a unique variance. This simplicity is due to our restriction of the function q to be quadratic in elements of ϕ.

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

4.3 Hyperparameters and Sampling

We set $\kappa_{0,A} = 100; \gamma_{0,A} = |I| + 1$. Parameters η_0 and Λ_0 were set to a vector of all zeros and the identity matrix, respectively. Importantly, we leveraged survey data for μ_0, setting values for identities equal to their mean in the survey data and setting $\kappa_{0,S} = 300$. Where no survey data existed for an identity (3% of the cases), the prior was set empirically, by running a simple dictionary-based model over a random 1% of the training data. The parameter Σ_0 was the identity matrix and $\gamma_{0,S} = 3000$. The Gibbs sampler was run for 500 burn-in steps, as models converged quickly (little variation was observed after about 300 iterations). We took five samples for model evaluation, one every 100 iterations from the 500th-900th steps of the sampler.

5. MODEL VALIDATION

While in the present work our focus is largely applied, it is still useful to provide some form of validation that the model we develop a) correctly learns parameters of interest and b) estimates parameters that generalize to unseen data within the population. With respect to the former goal, the appendix of this article presents a simulation validation study. With respect to the latter goal, we focus here on a brief evaluation of how well our model is able to perform on one task of particular interest to ACT scholars, namely the prediction of how an individual is labeled in a particular situation.

While we cannot replicate this task exactly as ACT scholars generally conceive of it, we can construct a related task by evaluating our model on its ability to predict which identities will appear in tweets in a held-out test set. So, for example, if the tweet "All girls rule!" was in the test set, evaluation in this section is aimed at seeing how well the model can predict that the identity "girl" is the true identity to be placed in that tweet, relative to all possible identities that could fit. In order to validate our model, we train on the first (temporally) 85% of each user's tweets and retain the last 15% for testing purposes.

In addition to seeing how predictive our model is, we also develop six baseline approaches. The first two are structural ablations of our model - we make predictions using only the semantic portion (*Model - Semantic Only*) or only the affective portion of the model (*Model - Affective Only*). We also compare our model to two different semantic-only baselines. The first is a multinomial over all counts for all users (*Simple Semantic*), and the second is a Laplace-smoothed language model (*User Semantic*). The latter is

Model	Average Rank
Simple Semantic	54.78
User Semantic	42.52
Simple Affective	134.74
User Affective	127.73
Model - Semantic Only	37.53
Model - Affective Only	126.04
Model - Full	67.54

Table 1: Results on the evaluation task.

equivalent to a model where each user is defined by a multinomial distribution over identities with a symmetric Dirichlet prior.[7]

Finally, we also compare against two affect-only baselines. To develop these affective baselines, we first run all training data through the VADER sentiment analysis tool [32], which gives a continuous value on the interval [-1,1] for each tweet. We then compute, both for each user and overall, the average sentiment of all tweets in which each identity occurred. This value serves as the affective stereotype for the identity. For each identity in each test tweet, we compute the sentiment score for the tweet with that identity removed from the text.[8] We then compare this sentiment score to either to the vector of averages for the user (the *User Affective* baseline) or to the average over all users (the *Simple Affective* baseline). For the user baseline, we use a simple fall-back model where the overall (simple) sentiment for an identity is used if the user has no tweets about a particular identity. These baseline models provide rankings based on which identities have affective stereotypes closest to the sentiment score for that test tweet. The affect-only baselines are similar to what is used in many applied sentiment analysis works, where keyword-based or tweet-level sentiment approaches are used to infer sentiment of concepts (e.g. [35, 13, 64]). They are therefore useful insofar as they can (in)validate our approach as a tool for applied, theory-driven research of affective stereotype on Twitter.

As an outcome metric, we measure for each model the *average rank* of the correct identity across all test tweets. For a single identity in a single test tweet, rank is determined by ordering all identities by their likelihood of being the correct identity for the tweet and then taking the index of the actual identity in this ranking. Average rank is simply an average over these rankings for all test tweets. A lower average rank therefore means that the model consistently placed the true identity closer to the top of its ranking.

Table 1 shows results on the validation task, where results from our full model and ablations of it are averaged across five Gibbs samples.[9] The best performing model was the semantic-only ablation of our model, which outperforms the user-based semantic baseline by around 1%. Performance of the full model falls in between the semantic-only and affective-only models. Finally, of the affective-only models, the structural ablation of our model performs best, improving by approximately 1% over the user-based affective baseline.

To test stability of these results, we also ran the evaluation task on models trained on five random 85/15 train/test splits of the data. In all five splits, ordering of the models was as seen within Table 1, and performance improvements of ablations of our model over their counterpart baselines were at least as high as shown above. This suggests results in Table 1 are fairly stable within our population.

The most obvious implication of these results is that affective information detracts from the predictive abilities of our full model. This finding is in line with observations that even frequency of term usage in text easily trumps affective information in a prediction setting [35]. This problem is exacerbated by the quadratic nature of the deflection equation, which leads to heavy-tailed predictions from the affective portion of the model. The affective portion is thus very "sure" of its top 3-5 predictions, leaving these to dominate the posterior of the full model's predictions.

Future work is therefore necessary to create an approach combining semantic and affective stereotypes that is highly predictive. However, for the purposes of the present work, where we are largely interested in an interpretation of semantic and affective stereotypes, the performance improvements of ablations of our model over the relevant baselines (i.e. comparing semantic to semantic and affective to affective models) suggests that the parameters learned by our model are at least as representative of underlying patterns in the data as other applied approaches we might have employed. We can therefore have as much confidence in interpreting these results as we might from a related research design, with the benefit of our model having a direct theoretical connection to ACT.

One final point relating to our evaluation is our decision not to use "gold-standard", hand-coded data to evaluate the affective portion of the model. While computational linguists routinely evaluate sentiment models on gold-standard datasets, we take the sociological perspective of Dimmagio [19], who argues that a true understanding of the affective meaning of a concept is not easily captured via human judgement after the fact. Further, in our case, there is little reason to believe that user stereotypes from this particular sample match our ACT survey data (in fact, we show evidence against this); we therefore do not use it or any annotator's judgements as a validation tool.

6. RESULTS

In this section we consider the semantic and affective stereotypes of the EG/MB population. We first seek confirmation that identity alignment to institutional structures impacts the network structure of semantic stereotypes. Finding this to be the case, we then consider how identities permeate the boundaries of these institutions and what this tells us about the semantic stereotypes of the EG/MB population. Finally, we look at how observations we make about the semantic stereotypes of the EG/MB population do or do carry over to semantic stereotypes estimated from a more recent, random sample of Twitter users.

We then move to a study of the affective stereotypes of the EG/MB population. Here, we leverage structure we observe in semantic stereotypes to hone in on identities that may have been particularly salient and thus possibly susceptible to changes in affective meaning. All results below are given for model parameters at one model sample (900th Gibbs sample); results are nearly identical at all other samples considered.

6.1 Semantic Stereotypes

Semantic stereotypes are captured in our model in the parameter matrix Λ. We follow prior work [7] and use the Graphical LASSO [24, 66]to sparsify Λ to better visualize and interpret sub-structures of this matrix. Figure 2 shows a network representation of Λ after applying a moderate level of sparsification, as defined by a parameter traditionally called λ. In the figure, nodes are identities and are colored by their cluster in the network as determined by the

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

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

[9]We would, in theory, take the average over many Gibbs samples. In practice, as with other researchers [46, 14], we find that variance across samples on the metric is low and therefore average over only a few.

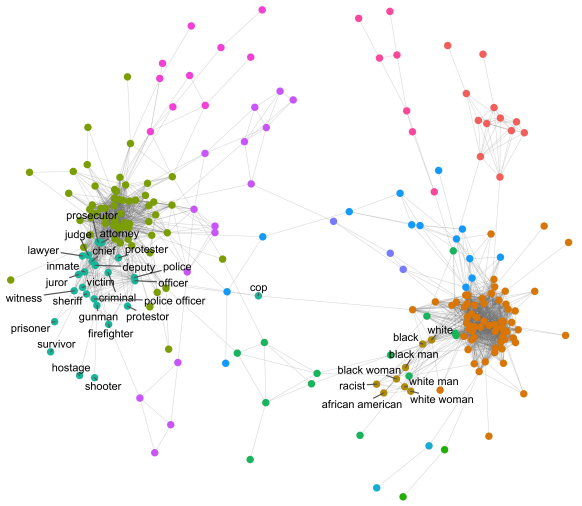

Figure 2: Network diagram of semantic stereotypes as estimated by the model parameter Λ, where we sparsify using $\lambda = .45$. Isolates are removed from the image. Labeled identities represent those of interest to our population, as described in the text.

Louvain method [8]. Links represent (only) positive semantic associations between two identities.

Table 1 displays the identities in each cluster extracted from the network displayed in Figure 2, along with a name for that cluster that we provided based on manual inspection. Many of the clusters in this network align with institutional settings uncovered in prior work [29, 35] (e.g. religion, medical, education, sports and deviant identities). There are also clusters that seem to blend two institutions, such as "Art/Music", "Religion/War" and "Legal/Protest". Finally, the "formal" and "informal" clusters combine identities tied to several different institutions. These clusters contain identities loosely relevant to an "informal", social context (daughter, idiot, guy, friend) and those we might use in a more news-oriented or formal discussion (e.g. republican, lawyer, priest), respectively. These two clusters are due in large part not to any underlying sociological factors but rather to the varying use of Twitter as both a platform for social interaction and information spread [63].

Notably, if we increase the sparsity parameter, we do find that clusters combining multiple institutions in Figure 2 seem to split further into clusters aligned with institutional settings. When we set $\lambda = .6$, for example, the Religion/War cluster splits into distinct clusters of religion (e.g. a cluster of pastor, Baptist, preacher) and military service cluster (e.g. vet and veteran). Similarly, the "formal" cluster partitions into several institutions, including one related to politics and one to business.

Results thus provide evidence semantic stereotypes connect identities into clusters aligned with institutional settings. However, this evidence must be qualified by the fact that both the structure of Figure 2 and the varying level of alignment to these institutional clusters at low levels of sparsification suggest that boundaries between these institutional settings are fuzzy. These fuzzy boundaries are created as individuals in our population move across institutional settings.

Figure 2 further shows that some boundaries are "fuzzier" than others. For example, the boundary between formal identities and identities in the Legal/Protest institutions is very fuzzy - individuals in the EG/MB population that discussed news-oriented topics like

Name	Identities
Informal	h$e, dad, ni$$a, best friend, girl, grandma, boyfriend, chick, bf, aunt, boy, baby, sibling, cousin, daddy, bro, friend, gangster, mom, dude, ...
Formal	democrat, conservative, candidate, senator, journalist, republican, governor, taxpayer, editor, liberal, politician, lawmaker, secretary, Muslim, activist, president, CEO, Hispanic, author, citizen, ...
Education	freshman, junior, academic, grad, college student, professor, principal, scholar, sophomore, student, cheerleader, teacher, intellectual
Sports	coach, player, fan, athlete, announcer, qb, pitcher, patriot, champion, teammate, winner, runner
Legal/Protest	**officer, attorney, police, deputy, judge, lawyer, chief, juror, gunman, protester, inmate, prosecutor, survivor, cop, firefighter, criminal, victim, police officer, sheriff, hostage, protestor, witness, prisoner, shooter**
Religion/War	pastor, vet, baptist, arab, civilian, Israeli, atheist, Christian, Jew, believer, marine, soldier, minister, preacher, veteran
Race	**black, black woman, white man, black man, white, African American, racist, white woman**
Art/Music	actor, musician, artist, actress, singer, celebrity, comedian
Sexuality	gay, lesbian, homosexual
Deviant	a$$hole, idiot, b$$tard, hypocrite, h$$ker, innocent, loser, liar, coward, murderer, killer, moron, punk, jerk
Religion	Catholic, pope, priest
Medical	nurse, doctor, patient
Tech	hacker, customer, designer, client, engineer, Russian, user, manager, photographer, spy
Hipster	hipster, geek, nerd
Asian	Asian, Chinese, Japanese

Table 2: Clusters of identities determined by applying the Louvain method to the network in Figure 2. Up to twenty (random) identities are shown for each cluster, and each cluster is given a name based on manual evaluation. Bolded clusters represent those particularly relevant to our population. Note that $ are used in place of letters for particular words.

politics also seem to have used legal and protest-based identities to discuss the Garner and Brown tragedies. In contrast, the boundary between the Race and Legal/Protest clusters is quite distinct - in fact, semantic relationships between identities within these clusters were non-existent.

The most interesting takeaway from Figure 2 is therefore not from the clustering of identities into institutions, as previous work from ACT scholars has focused on, but rather from the structure of boundaries across these institutions. In particular, we were most surprised to find that few interrelations existed between identities aligned with legal/protest institutions and more racialized identities. Even at lower levels of regularization ($\lambda = .3$), limited semantic connections emerge between the two clusters. At this level of regularization, the identity "cop" is found to be associated with nearly all of the racialized identities and the identity "racist" links to the identities "criminal", "police" and "cop".

These findings suggest that individuals who focused on the legal proceedings and/or protests surrounding the tragedies largely attempted to deracialize these events, whereas those focusing on race tended to focus on race alone. When these foci of discussion did cross, it was either with an intention of relating police, cops and criminals (broadly defined) to racists, or via the connection of racialized identities to cops specifically (rather than to, e.g., the jury or the protesters).

An important question is the extent to which these findings are unique to the EG/MB population. Intuitively, we would expect that

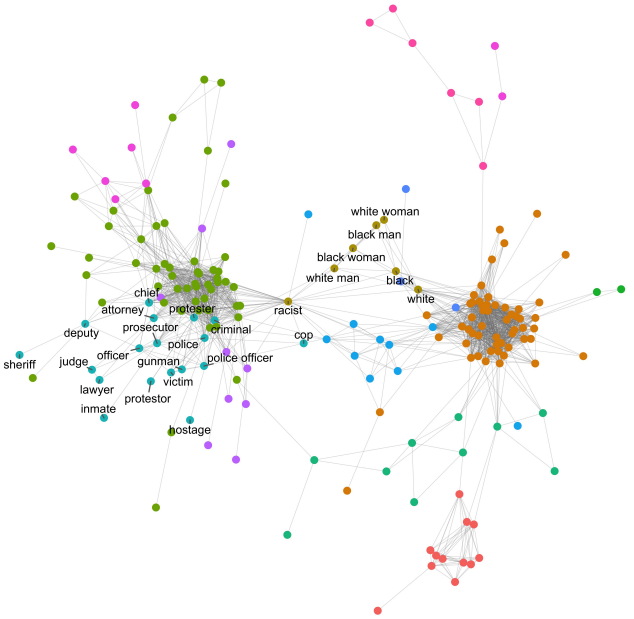

Figure 3: Network diagram of semantic stereotypes in a dataset of 10K random users. Colors for nodes are the same as those used in Figure 2 (but note the links have changed).

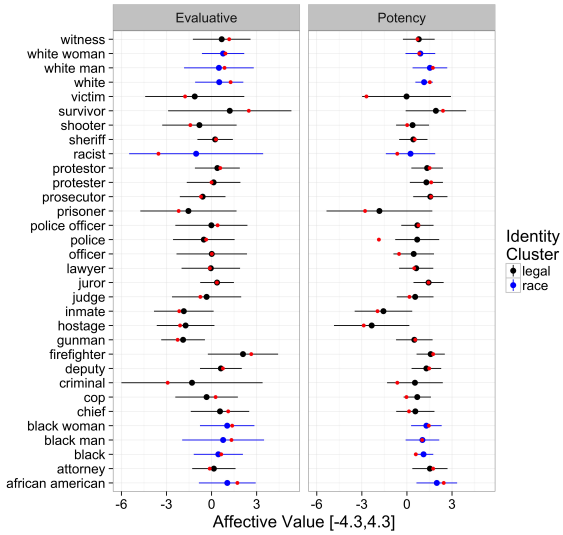

Figure 4: Measurements of the evaluative and potency dimensions (in the two distinct subplots, titles in grey) for all identities in the legal/protest and race clusters in Figure 2/Table 2. Black error bars represent $\mu \pm \sigma$ for the given affective dimension of the given identity, and red dots represent μ_0.

the clustering of identities into institutional settings should be relatively stable and thus extend beyond the EG/MB population, but that our observations about the Race and Legal/Protest institutions may be less stable. To explore this, we ran our model on a dataset of 10K users drawn randomly from the Streaming API who fit the characteristics of users in our study described above (i.e. less than 50K followers, etc.). Tweets from these users included messages through July of 2016, thus representing both a distinct population and a distinct period of study from the EG/MB population.

Figure 3 shows the resulting semantic stereotypes of these random users under sparsification conditions chosen so that the network displayed had similar size and density to the network in Figure 2 ($\lambda = .38$). The nodes in Figure 3 are colored by the *clustering of Figure 2* (i.e. clusters align with those listed in Table 2 and not based on a new clustering of the network in Figure 3). Both a visual inspection of Figure 3 and consideration of the high level of assortativity (.71) of the network based on the clusters from the EG/MB population provide evidence that the same institutional structures important to the EG/MB population are relevant in this random population as well. Further, the boundaries between these institutions are also consistent - individuals thus seem to discuss (or exist within) a fairly limited combination of institutional settings.

However, it is interesting to note that one (and perhaps, the only) place where these boundaries do change significantly surrounds a movement of the identity "racist" from the exterior of the semantic network in Figure 2 to the core of the network in Figure 3. The direct neighbors of the racist identity also change dramatically - in the Eric Garner/Michael Brown data, the only neighbors of the identity "racist" are other racialized identities - white, white man, black, black woman, black man, and white woman. In the randomized data, the top six identities correlated with racist are Muslim, terrorist, white man, rapist, hypocrite, and idiot. Undoubtably, as attention has shifted from the Garner and Brown tragedies to the tragedy of Donald Trump's presidential campaign, the semantic associations arising from ideas of racism have been altered.

In sum, we find that the strongest, most prevalent and most stable semantic stereotypes connect identities within the same institutional settings. However, the large shift in the position of racist when we considered a random sample of more recent Twitter users suggests that these institutional structures and their boundaries are not impervious to change. Further, we see that this story emerges not from simply enumerating clusters of identities but from exploring the "fuzziness" of institutional boundaries. These observations present a compelling reason for Affect Control scholars to reposition their focus from defining clusters of identities to moving towards the more network-based representation of semantic stereotypes put forward by cognitive psychologists.

6.2 Affective Stereotypes

From our exploration of semantic stereotypes above, we observed two clusters of identities - the legal/protest cluster and the race cluster- that were particularly relevant to the EG/MB population. Given what we know from prior work in ACT [56], the identities within these clusters may be particularly susceptible to systematic differences in affective stereotypes from the broader American public. To look for such differences, we can compare model estimates of stereotypes in the Twitter data (μ) to their nationally-representative, survey-based priors (μ_0).

Figure 4 displays the two most interpretable dimensions of affective stereotypes in ACT, the evaluative and potency dimensions [50], for all identities in the Race and Legal/Protest clusters in Table 2. From Figure 4, we see that the only case in which the survey prior differs widely from the distribution of affective stereotypes in the EG/MB population is on the potency dimension of the police identity. Specifically, the police identity as perceived by the individuals in the EG/MB population is significantly more powerful than we would expect given the survey data.

From Figure 2, we know that the police identity was most likely to be discussed by individuals who also discussed other legal identities, as well as identities associated with "formal" discussions of topics like politics and business. Consequently, results may be biased towards views of a subset of our users. Regardless of this

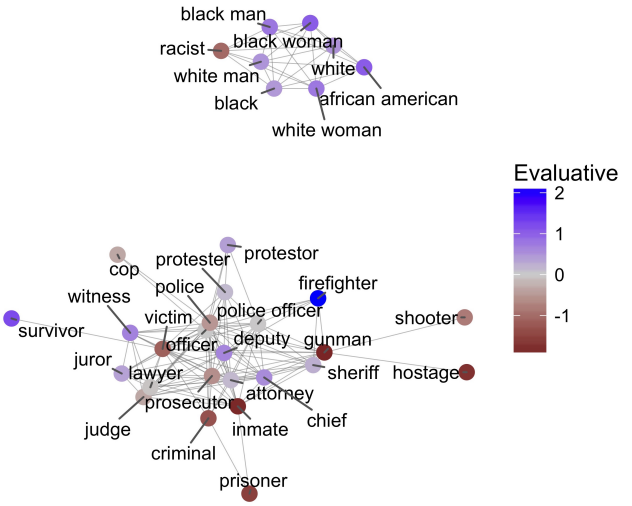

Figure 5: Network views of the two identity clusters of particular relevance to our population - the legal/protest cluster and the race cluster. Links are the same as those visualized in Figure 2 and identities are colored by their evaluative meaning.

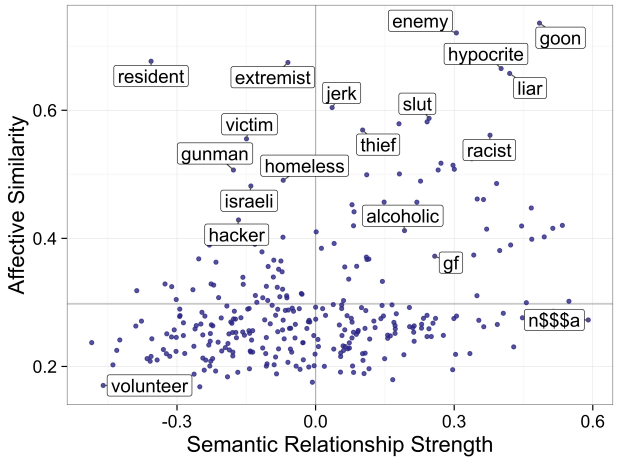

Figure 6: The x-axis gives each identity's measured semantic relationship to thug, the y-axis gives the identity's affective similarity to thug, computed as the one over the square root of the unnormalized Euclidean distance between measured EPA profiles. Only outlier points are labeled. Grey lines on the x- and y-axes represent a null semantic relation and the mean affective similarity, respectively. A point is shown for all identities except for thug. Note that $ are used in place of letters for particular words.

point, what can be said is that the affective dimension of the police identity that is important to these individuals the power dimension and not the evaluate ("good/bad") dimension. This importance given by the EG/MB population to power accurately captures a relevant factor of the debate on policing in America on a dimension of affect not traditionally studied, suggesting future analyses would do well to think of affective meanings of identities beyond the simple "good/bad" approach taken in most existent sentiment analysis approaches available to scholars.

7. DISCUSSION - COMBINING AFFECTIVE AND SEMANTIC STEREOTYPES

To this point, we have focused on interpreting affective and semantic stereotypes independently of one another. In this discussion section, we consider ways in which combining these two kinds of stereotype more explicitly might help us to better understand both the structure of particular institutions and the meaning of particular identities.

To the extent that institutions can be isolated from each other, combining semantic and affective stereotypes might help to identify generalizable role relationships. As an example, Figure 5 presents the identities from the legal/protest and race institutions from Figure 2, colored by their affective meaning on the evaluative dimension. The race institution is a near-clique - with the exception of "racist" and "African American", all identities within the institution are semantically related to each other. Further, only the identity "racist" has a negative evaluation. Within the race institution, Figure 5 thus suggests that as viewed by the EG/MB population, the racist identity takes on a very clear role as the deviant identity within the institution. In contrast to the race institution, the legal/protest institution displays a core-periphery structure and a high degree of assortativity in evaluative meaning. Within this institution, a variety of role relations thus might be found. For example, inmates and firefighters define the polar ends of the evaluative spectrum. Future work might consider analogous pairs across a variety of institutions, or consider how these identities may act as affective anchor points in particular settings.

Progress on these questions will require further theoretical work on the meaning and generalizability of role relationships defined by both affective and semantic stereotypes. However, the stereotypes extracted by our model can be used in a more concrete way to better understand individual identities. One identity of interest to the EG/MB population we have not yet discussed is the identity thug, which appeared in the "informal" cluster in Table 2. As John McWhorter notes in lieu of media coverage of the Freddie Gray protests in Baltimore, "...thug today is a nominally polite way of using the N-word...It is a sly way of saying there go those black people ruining things again".[10] McWhorter argues that while the news media was using the identity thug to distinguish protesters who turned violent, the term was really being used in lieu of more racialized identities.

If this connection were to be present in wider swaths American culture, we should expect to see it emerge in social media data. One method we can use to confirm this is to look at existing data capturing meaning in large quantities of such data. Sure enough, the top five most similar words to thug that are not alternative spellings of the word (e.g. thugs) in the publicly available 200-dimensional GloVe word embeddings [48] trained on a large amount of Twitter data are gangsta, ni$$a[11], goon, lil and homie. Within these embeddings, there is an obvious blend of both affectively negative identities (goon) and identities that either refer to or are often used by the African American community.

Results from our model also retain this connection between thug and identities related to the African American community. However, as Figure 6 shows, by separating out semantic stereotypes from affective stereotypes, our model allows one to more readily discern nuanced relationships between thug and other identities. For example, our model agrees with the GloVe embeddings in that thugs are "culturally synonymous" with goons - that is, these two identities have similar affective meanings and strong semantic sim-

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

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

ilarity. However, Figure 6 also shows that our model differentiates thugs and "ni$$a" along the affective dimension, suggesting that while similar people refer to these two identities frequently, it is done so in different emotional contexts.

Affective meaning may in this way be a useful tool to help differentiate between true synonymy and simple association - words people feel differently about may be used frequently by the same set of people but are unlikely to refer to the same underlying concept. More generally, this observation makes the case for future work on word embeddings that focuses on understanding their biases [9, 11], the meaning of particular embedding dimensions [41] and on incorporating affective meanings in intelligent ways [58].

8. CONCLUSION

The core contributions of the present work are, first, the development of a publicly available method to rapidly infer the affective and semantic stereotypes of a population of Twitter users, and second, the application of this method to a population of Twitter users who actively discussed the Eric Garner and Michael Brown tragedies.

With respect to the former contribution, our method provides an important new tool, both theoretically and empirically, for researchers of Affect Control Theory. Empirically, we present a new method to infer stereotypes of a population with tunable priors allowing for arbitrary adherence to existing survey data. Perhaps more importantly, our work challenges the theoretical assumption that semantic associations can simply be written off to static institutional settings. Instead, an explicit mathematical modeling of semantic stereotypes as a network shows that while institutional structures are important in determining semantic stereotypes, important structure also exists *between* institutions that can be useful in understanding the population of interest. Further, network models of semantic stereotypes provide a stronger connection to models of stereotype faithful to cognition.

With respect to the second contribution of our work, we find that affective stereotypes of the EG/MB population are for the most part stable as compared to nationally representative survey data. However, we do find evidence that this population views the police as being significantly more powerful than this survey data would suggest, implying that the core struggle of the protests resulting from these tragedies may be most relevant to the power police hold over communities of color, rather than the extent to which police are "good" or "bad". We also find evidence in the structure of the semantic stereotypes we measured that discussions of racialized identities and legal/protest identities were carried out by individuals situated in relatively disparate institutionalized settings. One straightforward interpretation of this finding is that those in more racialized social contexts tended to "see" and thus discuss race as it was perceived to be relevant to the tragedies, whereas those in less racialized contexts "saw", or at least chose to discuss, these events only in terms of legal proceedings and ensuing protests. These findings speak to the importance of how we select identities for ourselves and others, both in our online expressions and in our everyday lives.

In considering the findings presented here, it is important to realize that the Twitter data that we use is biased in important ways [42, 52, 60], and thus our results, where not explicitly stated otherwise, should not be seen as generalizing even to the entirety of Twitter invested in these tragedies, let alone to more general social settings. Further, while our model was not geared towards prediction, validation of the model suggests room for improvement in how well parameter estimates from our model generalize to unseen data. Consequently, estimated parameters may be biased in ways that are not immediately clear to us at present, particularly within the affective portion of the model.

Future work might well try to address these limitations, finding ways to blend more predictive models of text (i.e. neural models) with models such as ours that retain an element of explainability and adherence to social theory. As noted above, future work might also do well to automatically capture variations within the population of study itself in terms of its semantic and/or affective stereotypes. We look forward future work combining NLP and sociocognitive theory on this matter.

9. ACKNOWLEDGEMENTS

We would like to thank the reviewers for this paper, who offered several extremely helpful critiques that tremendously improved the article. We would also like to thank Jon Morgan for his candid-as-always insights into the problem, as well as the first author's committee members for their valuable suggestions. Support was provided, in part, by the Office of Naval Research (ONR) through a MURI N00014081186 on adversarial reasoning and the ONR through a MINERVA N000141310835 on State Stability.

10. REFERENCES

[1] Areej Ahothali and Jesse Hoey. 2015. Good News or Bad News: Using Affect Control Theory to Analyze Readers' Reaction Towards News Articles. ACL (2015).

[2] John R Anderson. 2007. *How can the human mind occur in the physical universe?* Oxford University Press, Oxford [etc.].

[3] John R. Anderson, Michael Matessa, and Christian Lebiere. 1997. ACT-R: A theory of higher level cognition and its relation to visual attention. *Human-Computer Interaction* 12, 4 (1997), 439–462.

[4] David Bamman, Ted Underwood, and Noah A. Smith. 2014. A bayesian mixed effects model of literary character. In *Proceedings of the 52st Annual Meeting of the Association for Computational Linguistics (ACL'14)*.

[5] Marianne Bertrand and Sendhil Mullainathan. 2003. *Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination*. Technical Report.

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

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

[8] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (2008), P10008.

[9] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *arXiv preprint arXiv:1607.06520* (2016).

[10] Eduardo Bonilla-Silva. 2013. *Racism without Racists: Color-Blind Racism and the Persistence of Racial Inequality in America* (4 edition ed.). Rowman & Littlefield Publishers, Lanham.

[11] Aylin Caliskan-Islam, Joanna J. Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from

language corpora necessarily contain human biases. (Aug. 2016). arXiv: 1608.07187.

[12] Kathleen Carley and Michael Palmquist. 1992. Extracting, representing, and analyzing mental models. *Social forces* 70, 3 (1992), 601–636.

[13] Nathanael Chambers, Victor Bowen, Ethan Genco, Xisen Tian, Eric Young, Ganesh Harihara, and Eugene Yang. 2013. Identifying Political Sentiment between Nation States with Social Media. *Computational Linguistics* 39 (2013), 4.

[14] Jianfei Chen, Jun Zhu, Zi Wang, Xun Zheng, and Bo Zhang. 2013. Scalable inference for logistic-normal topic models. In *Advances in Neural Information Processing Systems*. 2445–2453.

[15] Lu Chen, Wenbo Wang, Meenakshi Nagarajan, Shaojun Wang, and Amit P. Sheth. 2012. Extracting Diverse Sentiment Expressions with Target-Dependent Polarity from Twitter.. In *ICWSM*.

[16] Allan M. Collins and Elizabeth F. Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological review* 82, 6 (1975), 407.

[17] Lauren D. Davenport. 2016. The Role of Gender, Class, and Religion in Biracial Americans' Racial Labeling Decisions. *Amer. J. Sociology* 81, 1 (2016), 57–84.

[18] James Deese. 1966. *The structure of associations in language and thought*. Johns Hopkins University Press.

[19] Paul DiMaggio. 2015. Adapting computational text analysis to social science (and vice versa). *Big Data & Society* 2, 2 (Dec. 2015), 2053951715602908.

[20] Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*. 49–54.

[21] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2014. Diffusion of Lexical Change in Social Media. *PloS one* 9, 11 (2014), e113114.

[22] Ethan Fast, Tina Vachovsky, and Michael S. Bernstein. 2016. Shirtless and Dangerous: Quantifying Linguistic Signals of Gender Bias in an Online Fiction Writing Community. *arXiv preprint arXiv:1603.08832* (2016).

[23] Jonathan B. Freeman and Nalini Ambady. 2011. A dynamic interactive theory of person construal. *Psychological review* 118, 2 (2011), 247.

[24] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9, 3 (2008), 432–441.

[25] David R. Heise. 1979. *Understanding events: Affect and the construction of social action*. CUP Archive.

[26] David R. Heise. 1987. Affect control theory: Concepts and model. *The Journal of Mathematical Sociology* 13, 1-2 (Dec. 1987), 1–33.

[27] David R. Heise. 2007. *Expressive Order*. Springer.

[28] David R. Heise. 2010. *Surveying cultures: Discovering shared conceptions and sentiments*. John Wiley & Sons.

[29] David R. Heise and Neil J. MacKinnon. 2010. *Self, identity, and social institutions*. Palgrave Macmillan.

[30] Jesse Hoey and Tobias Schröder. 2015. Bayesian affect control theory of self. In *Proc. of the AAAI Conference on Artificial Intelligence*. bibtex: hoey_bayesian_2015.

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

[32] C. J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAAI Conference on Weblogs and Social Media*.

[33] Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 151–160.

[34] Kenneth Joseph, Wei Wei, Matthew Benigni, and Kathleen M Carley. 2016b. Inferring affective meaning from text using Affect Control Theory and a probabilistic graphical model. *Journal of Mathematical Sociology* (2016).

[35] Kenneth Joseph, Wei Wei, and Kathleen M Carley. 2016a. Exploring patterns of identity usage in tweets: a new problem, solution and case study. In *WWW 2016*.

[36] Svetlana Kiritchenko, Xiaodan Zhu, and Saif M. Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research* (2014), 723–762.

[37] Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. 2014. A dependency parser for tweets. In *Proc. of the Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, to appear*.

[38] Petra Kralj Novak, Jasmina Smailovic, Borut Sluban, and Igor Mozetic. 2015. Sentiment of Emojis. *arXiv preprint arXiv:1509.07761* (2015).

[39] Ziva Kunda and Paul Thagard. 1996. Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review* 103, 2 (1996), 284–308.

[40] Kun-Lin Liu, Wu-Jun Li, and Minyi Guo. 2012. Emoticon smoothed language models for twitter sentiment analysis. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

[41] Quan Liu, Hui Jiang, Si Wei, Zhen-Hua Ling, and Yu Hu. 2015. Learning semantic word embeddings based on ordinal knowledge constraints. *Proceedings of ACL, Beijing, China* (2015).

[42] Momin M. Malik, Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer. 2015. Population Bias in Geotagged Tweets. In *Ninth International AAAI Conference on Web and Social Media*.

[43] George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes* 6, 1 (1991), 1–28.

[44] Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*. Association for Computational Linguistics, 26–34.

[45] Jonathan H. Morgan, Kimberly B. Rogers, and Mao Hu. 2015. Distinguishing Normative Processes from Noise: A Comparison of Four Approaches to Modeling Impressions of Social Events. In Submission (2015).

[46] Brendan O'Connor, Brandon M. Stewart, and Noah A. Smith. 2013. Learning to Extract International Relations from Political Context.. In *ACL (1)*. 1094–1104.

[47] James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71 (2001), 2001.

[48] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)* 12 (2014), 1532–1543.

[49] Soujanya Poria, Erik Cambria, GrÃl'goire Winterstein, and Guang-Bin Huang. 2014. Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems* 69 (2014), 45–63.

[50] Kimberly B. Rogers, Tobias Schröder, and Wolfgang Scholl. 2013. The Affective Structure of Stereotype Content Behavior and Emotion in Intergroup Context. *Social Psychology Quarterly* 76, 2 (2013), 125–150.

[51] Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M. Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. Semeval-2015 task 10: Sentiment analysis in twitter. In *Proc. of the 9th International Workshop on Semantic Evaluation, SemEval*.

[52] Derek Ruths and Jürgen Pfeffer. 2014. Social media for large studies of behavior. *Science* 346, 6213 (2014), 1063–1064.

[53] Hassan Saif, Yulan He, Miriam Fernandez, and Harith Alani. 2015. Contextual semantics for sentiment analysis of Twitter. *Information Processing & Management* (2015).

[54] Tobias Schr oder, Jesse Hoey, and Kimberly B. Rogers. 2017. Modeling dynamic identities and uncertainty in social interactions: Bayesian affect control theory. *Am. Soc. Rev* (2017).

[55] Tobias Schröder and Paul Thagard. 2014. Priming: Constraint satisfaction and interactive competition. *Understanding Priming Effects in Social Psychology* (2014), 157.

[56] Lynn Smith-Lovin and William Douglas. 1992. An affect control analysis of two religious subcultures. *Social perspectives on emotion* 1 (1992), 217–47.

[57] L. Smith-Lovin and Dawn T. Robinson. 2015. Interpreting and Responding to Events in Arabic Culture. *Final Report to Office of Naval Research, Grant N00014-09-1-0556* (2015).

[58] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, Vol. 1. 1555–1565.

[59] Lisa Thomas and David R. Heise. 1995. Mining Error Variance and hitting pay-dirt: Discovering systematic variation in social sentiments. *The Sociological Quarterly* 36, 2 (1995), 425–439.

[60] Zeynep Tufekci. 2014. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. In *ICWSM '14: Proc. of the 8th International AAAI Conference on Weblogs and Social Media.*

[61] Duy-Tin Vo and Yue Zhang. 2015. Target-dependent twitter sentiment classification with rich automatic features. In *Proc. of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*. 1347–1353.

[62] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods* 45, 4 (2013), 1191–1207.

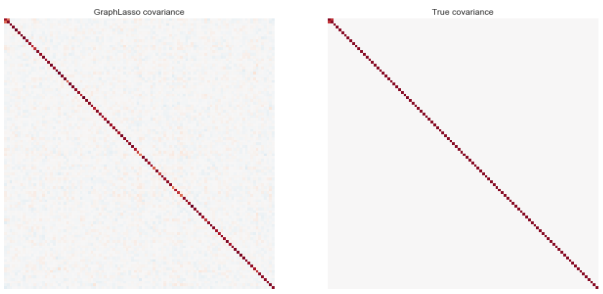

Figure 7: On the left, the learned covariance matrix by our model (after sparsification with the graphical LASSO). On the right, the true simulated covariance matrix

[63] Lei Yang, Tao Sun, Ming Zhang, and Qiaozhu Mei. 2012. We know what @you #tag: does the dual role affect hashtag adoption?. In *Proceedings of the 21st international conference on World Wide Web (WWW '12)*. ACM, New York, NY, USA, 261–270.

[64] Amy X. Zhang and Scott Counts. 2016. Gender and Ideology in the Spread of Anti-Abortion Policy. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 3378–3389.

[65] Meishan Zhang, Yue Zhang, and Duy-Tin Vo. 2015. Neural networks for open domain targeted sentiment. In *Conference on Empirical Methods in Natural Language Processing*.

[66] Tuo Zhao, Han Liu, Kathryn Roeder, John Lafferty, and Larry Wasserman. 2012. The huge package for high-dimensional undirected graph estimation in r. *The Journal of Machine Learning Research* 13, 1 (2012), 1059–1062.

APPENDIX

The goal of this appendix is to ensure that our model can correctly infer parameters generated from known distributions. Like nearly all Bayesian models of text, the generative story of our model first requires that we draw a number of documents for each user. In social media, the number of messages sent per user is heavy-tailed. We found that users who had less than 175 tweets led to unstable inference, we thus generate the number of "tweets" per user from a log-normal distribution (a well-known heavy-tailed distribution) and ensure that each simulated user (as in the real model) has at least 175 tweets. As with all simulations presented here, we run with 5000 "users" and 100 "identities", where quotes are used to indicate that these are simulated. In other words, the number of tweets per user, N_u, is drawn for our simulation from the following distribution: $N_u \sim \exp(\mathcal{N}(4.86, 1.22)) + 175$.

For the semantic portion of the model, we are largely interested in the parameter Λ - the association matrix. We model a single pair of correlated variables - we set $\Lambda_{0,1} = \Lambda_{1,0} = .8$ and restrict the rest of Λ to the identity matrix. We then draw activation scores (ν) for the simulated data for each identity from $\mathcal{N}(5, 1)$, and finally, we simulate data for each user by drawing their activation values from $\phi \sim \mathcal{N}(\nu, \Lambda)$.

This essentially models the count of each identity overall as a log-normal distribution as well, which is reasonable for real-world data also. We run the semantic portion of the model for fifty iterations in order to learn parameters, initializing the model with the same parameters as in the text. The left plot in Figure 7 shows the estimated covariance matrix from a Graphical LASSO of the

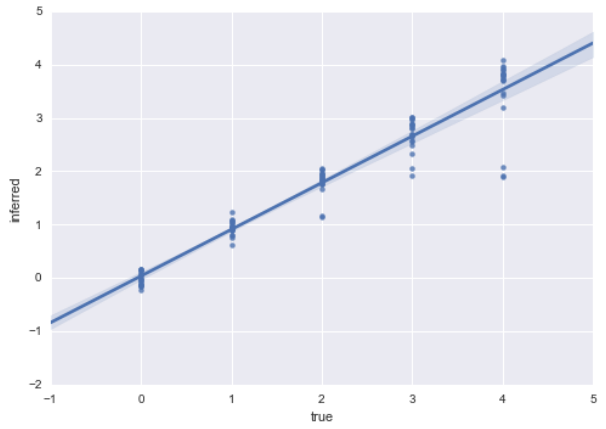

Figure 8: On the x-axis, the true value of each index of μ used to simulate the data - on the y-axis, the inferred value

estimated Λ (using cross-validation to set the degree of regularization). The right plot in Figure 7 shows the true covariance matrix we specified via the simulation. Results show that the model indeed learns the correct parameterization.

For the sentiment portion of the model, we first generate μ by setting the value of each index of μ to be the index modulo 5 - thus, for example, the 6th entry of μ is given a value of 1. The parameter Σ is set to the identity matrix. We then simulate the creation of tweets for users by assuming that each tweet contains a single identity (drawn with probability proportional to ϕ for that user) with a single constraint - for simplicity, let us assume the constraint is only on the *e*valuative dimension.

We initialize the model by setting all indices of μ to zero and Σ to the identity matrix, setting $\kappa = 100, \nu_0 = 1000$ (slightly lower than in the text to address the fact we are simulating fewer users). We run the model for 20 iterations and focus on whether or not the model is able to infer the true values of μ. Figure 8 shows that the model is able to infer the true parameterization in almost all cases, but that the prior does impact the model at higher values. As discussed in the text, this means that model estimates are likely slightly over-biased towards retaining their value in the survey data, which we feel is acceptable given the way the survey data was collected.