The Identity Labeling Problem and a Corresponding "Solution"

Kenneth Joseph Northeastern, Harvard, & SUNY Buffalo

Jonathan Howard Morgan Duke University

January 11, 2018

Abstract

The present work introduces and mathematically formalizes the *identity labeling problem* - given an individual in a social situation (the "labelee"), can we predict what identity(ies) he or she will be labeled with by someone else (the "labeler")? We argue that existing predictive models of identity labeling are theoretically incomplete and provide survey results that confirm this intuition. We then introduce a novel approach to the identity labeling problem, which we call *Latent Cognitive Social Spaces (LCSS)*. We show that LCSS is a better predictor of identity labeling in survey data than previous models and discuss its theoretical implications for linking the definition of the situation to theories of identity in action.

1 Introduction

Identities are the labels we use to denote our social roles, categories and group memberships (Tajfel and Turner, 1979; Smith-Lovin, 2007). It is generally accepted that the way we label ourselves, the way we label others, and the ways they label us all impact our behavior. Explaining how identity has such effects is the purview of numerous social theories focusing on how behavior changes within pre-defined social contexts and pre-assigned identities. Elements of the social context and pre-determined identities therefore act as scope conditions for such theories (e.g., expectations states theory, identity control theory). Outside of the scope of these "behavioral theories," though, is a model of how social actors decide which identity(ies) to apply to themselves and others in the first place. The present work focuses on a better understanding of this phenomena. In doing so, we hope improve understandings of how identity impacts behavior as well.

Specifically, we here study the following question - given an individual in a social situation (the "labelee"), can we predict what identity(ies) he or she will be labeled with by someone else (the "labeler")? We will refer to this as the *identity labeling problem*. The identity labeling problem is a foundational and unresolved component of the definition of the situation (Thomas and Thomas, 1928), the process by which a social actor internalizes and organizes a social

situation in her head in order to engage in the "self-determined lines of action and interaction" (Ball, 1972, p. 63) studied by behavioral theories like those noted above.

For a theory to make predictions about the identity labeling problem, two core sub-problems must be addressed. First, the theory must address dimensions of salient information the labeler will use to identify the labelee. Second, the model must determine the identity(ies) that the labeler will believe best "fits" the labelee. With respect to what information is salient for identity labeling, an overwhelming amount of research suggests that visible features of individuals, such as their skin color, are almost universally leveraged to label others (see, e.g., Cikara and Van Bavel, 2014, for a neuroscience-oriented review). Labelers also derive salient features from the current social context - for example, doctors are more likely to be seen in hospitals than in schools (Heise and MacKinnon, 2010). Finally, labelers are likely to leverage behaviors that the labelee is engaging in to determine the labelee's identity (Heise, 1987).

In order to then "fit" salient information about the labelee (within a context) to a particular social identity, a model of identity labeling then requires a means of predicting how the labeler will match the observed information to all possible identities she could choose to apply. This matching process, in turn, requires a means of characterizing each potential identity in some "meaning space" of potentially salient dimensions, as well as a method of determining which of these dimensions is most important to the labeler at a given time.

Existing scholarship provides a wealth of knowledge about what information is important in the identity labeling process and how different identities "fit" particular bits of information. However, only a small subset of existing research proposes models that can make *predictions* about how a labeler will identify a labelee in a social situation. Without the ability to assess generalizable predictions of theoretical models of identity labeling, we run the risk of assuming the model holds in cases where it does not. This can have drastic impacts on our understanding of how identity impacts behavior - if our model of identity labeling misjudge which identities are being applied, then we can in turn misjudge, for example, effects of the impacts of different prejudices on employment (Gaddis, 2017).

The present work studies two of the only existing classes of mathematical models that, suitable defined, can make predictions about identity labeling—*Affect Control Theory* (Heise, 1987) and the parallel constraint satisfaction model of Freeman and Ambady (2011). Because these two models draw on vastly different literatures, we formalize the identity labeling problem mathematically and show how each of these models, appropriately defined, can be used to explain how individuals are labeled.

We then develop and carry out a survey designed to test the ability of these models to make predictions in a naive version of the identity labeling problem. We show that neither the Freeman and Ambady (2011) model, which focuses largely on "cognitive", or "semantic" forms of information, like visual traits and contextual cues, nor Affect Control Theory, which focuses solely on "affective" information, can make high-quality predictions of survey responses.

To address the shortcomings of existing predictive models of the identity labeling model, we propose a new predictive model, entitled *Latent Cognitive Social Spaces* (LCSS), which is the first we are aware of to incorporate salient

affective and cognitive information in the context of identity labeling. After describing the model, we provide evidence that LCSS is able to make more accurate predictions than the other two models. Because LCSS is derived from a model linking identity to behavior, it provides an important new bridge between models of the complementary processes of defining the social situation and behaving within it.

2 The Identity Labeling Problem

Let us use an example to further explicate the identity labeling problem. Imagine that you are in a courtroom in the U.S. and observe an individual with female-sex indicative traits who is lecturing a lawyer and wearing a robe. The identity labeling problem refers to the problem of predicting the identity(ies) you will select for this person (out of all possible identities you could choose). In this hypothetical situation, the most likely label is obvious; only judges wear legal robes in the U.S. However, any model used to actually predict this conclusion is not so easily contrived.

To do so, such a model must answer at least two underlying questions. First, what salient information might you use to make your decision? In general, three "kinds" of meanings have been used in prior work. First, *traits* of the individual to be labeled might be used. Here, for example, the fact that the individual is wearing a robe may be consequential. Second, *semantic associations* we derive from other labeling decisions we have already made can impact our decision. For example, labeling our setting as a "courtroom" and another individual in this setting as a "lawyer" would surely influence the scope of identities you are willing to consider. Finally, *affective meanings*, chiefly informed by social events, provide you with clues. Here, the fact that the to-be-labeled individual is lecturing a lawyer, an action implying a certain degree of power, may hone your choice of identity.

Second, how does this salient information fit the *meaning* of the possible identities? Characterizing this fit can be difficult; what is to be done when observed signals disagree? And how do we transfer salient information into particular theorized dimensions of identity meaning? For example, what affective meaning does the social act of lecturing convey, and what identities best match that affective meaning?

Predictive models for identity labeling therefore are, implicitly or explicitly, constructed in two parts. First, they define some kind of meaning space, M, in which identities and all relevant salient information within the social environment can be defined and/or translated into. Second, identity labeling models define some function, ϕ, that provides a quantification of how well it expects the labeler to believe each potential identity label "fits" the given information. By applying ϕ across all possible identities, models can determine the identity(ies) the labeler will most likely apply to the labelee.

In the present work, we show that existing predictive models of the identity labeling problem do not fully characterize M, the meaning space of identities and salient information. We show that this is the case by developing a survey where existing models of identity labeling cannot provide correct predictions. We then define a new M (and

consequently, a new ϕ) by combining previous models that allows us to predict the survey responses more accurately. We refer to this new model of identity labeling as Latent Cognitive Social Spaces (LCSS).

3 Related Work

As the aim of the present work is the development of a predictive model, we focus largely on work in sociology and cognitive psychology that develops quantitative models relevant to the identity labeling problem. We review sociological and cognitive literature separately, focusing both on how scholars quantitatively represent meanings of identities and how they develop predictive models of identity labeling.

3.1 Sociological and Social Psychological Models

The Stereotype Content Model (SCM) (Fiske et al., 2002) is one the few theories to pose a quantitative model of identity meaning. SCM theorizes that identity meanings are defined by a two-dimensional affective space. The axes of this space define the "warmth" and "competence" of identities. The affective meaning of identities along the warmth and competence dimensions are assumed to be universal to the extent that, when asked "what do you think people in general think about identity X", individuals within a particular "national culture" will give similar responses.

Like most sociological models relevant to the identity labeling problem, the SCM focuses on defining a meaning space for identities and then assessing consequences of those meanings once identities have been applied to individuals. In the mathematical language introduced above, then, the core focus is to define M and then to assess the impact of identity labeling once an (unknown) ϕ has been applied.

Recently, Rogers et al. (2013) compared the meaning model for identities of SCM to that of affect control theory (ACT). ACT is a sociological model of identity and action (Heise, 2007). It has three core tenets. First, ACT assumes an (affective) meaning space for identities and the behaviors they engage in similar to that proposed by the SCM. Like the SCM, ACT assumes that identity meanings are generally universal. In ACT, however, each identity and behavior is defined by a three-dimensional affective profile. The first dimension, the Evaluative dimension, specifies how "good" or "bad" an identity is. The Potency dimension specifies the powerfulness/powerlessness of the identity, and the Activity dimension defines the activeness/passiveness of the identity. Rogers et al. (2013) find that the warmth and competence dimensions of the SCM are highly correlated with ACT's evaluation and potency dimensions respectively.

The second core tenet of ACT is an assumption that during social situations, actors engage in social events, or actions that one individual takes towards another, based on culturally shared fundamental meanings of identities and behaviors. Finally, ACT assumes that the way we interpret and choose to engage in social events, and the way we (re)label participants in those events, is done in a way so as to maintain these fundamental meanings of identities and behaviors. This final portion of the model is often referred to as the control principle of the model.

ACT is traditionally used as a model that relies on predetermined identity labelings - scholars take as given two hypothesized identities and use the control principle to study likely behaviors between the two identities. However, ACT can also be used to infer the identity of an individual that engages in a social event when the identity of the other is known. For example, given that a teacher is instructing "someone", we can predict using ACT that the most likely identity of "someone" is "student." In this form of the identity labeling problem, the meaning space of identities, M, is EPA space and the salient information used to make labeling decisions can be defined simply as the information that a teacher is instructing the labelee. The function ϕ is defined by the parameters of an impression change equation, described in more detail below. This equation is used to determine how likely a given identity is to apply to an unlabeled individual within a social event.

This use of ACT provides the only quantitative sociological model we are aware of in which the identity labeling problem is formalized and generalizable predictions are created. It should be noted, however, that recent work on the theory has focused on addressing shortcomings to the original model, leading to slightly different theoretical tenets. Specifically, *BayesACT*, a probabilistic extension of the original ACT model (Hoey et al., 2013; Hoey and Schröder, 2015; Schröder et al., 2017), addresses the fact that ACT in its original form focuses only on point estimates of EPA profiles. In Section 5 we provide more details on the mathematics of ACT and BayesACT.

While (Bayes)ACT can therefore be used to address the identity labeling problem in one form, it was not developed for this purpose. Consequently, aspects of the theory make it less reliable as a predictive model for the identity labeling problem. These difficulties are made clear by the following example, where artificial agents in simulations of BayesACT revert to situations like the following, quoted from Schröder et al. (2017): "...both agents have developed the shared belief that one of them (agent A) is an "executioner" while the other (agent B) is a "great grandmother."

While *affectively*, this shared belief may make sense given the behaviors these agents engage in towards each other, intuition tell us that the identity pairing of "executioner" and "great grandmother" is unlikely in any real world scenario. Affect Control researchers have realized that such intuition derived from semantic, or cognitive, meanings of identities. To address this, ACT scholars have developed the concept of institutions - clusters of identities and behaviors that can be grouped together via institutional settings (Heise and MacKinnon, 2010). Institutions act as a "filter" for the identities available to us when labeling an individual. So, for example, if we know that an individual is in the "school" institutional setting, ACT assumes that we will restrict labels for that individual to identities within the "school" institution, like principle and student.

This understanding of semantic meanings, while addressing intuitive discrepancies, is not likely to be predictive of identity labeling for three reasons. First, semantic relationships between identities frequently cross institutional boundaries. For example, teachers (the "school" institution) often interact with parents (the "family" institution). Second, some semantic relationships are stronger than others. For example, the semantic association between "student" and "teacher" is probably stronger than the one between "student" and "school administrator." Finally, it is not clear

that there is always a two-step process by which we first rule out semantically unreasonable identities and then apply affective meanings. Indeed, results shown in the present work strongly suggest otherwise.

(Bayes)ACT thus gives an important conceptual model of how affective and semantic association meanings might be combined to make predictions for a specific form of the identity labeling problem. Further, it provides a grounding for how we can connect a model of identity labeling to the behaviors that these labelings produce. However, the precise mechanisms specified are not likely to be useful in making predictions for the identity labeling problem, as we desire here. In particular, the meaning space of ACT seems inadequate, as semantic representations of ACT are underspecified, and little is done to incorporate trait-based stimuli. Fortunately, other models for the identity labeling problem provide insights into the effects of these forms of meaning.

3.2 Cognitive Models

Models of the identity labeling problem in cognitive psychology are semantic in nature. As opposed to institutional semantics, however, cognitive models are instead relational. By this, we mean that parameterizations are defined by where links, or semantic relationships, do (or do not) exist between identities themselves and between identities and environmental cues. Here, we distinguish between two "kinds" of semantic relationships. First, links between features of individual to be labeled (e.g. skin color, age, etc.) and particular identities will be defined as *traits*. Second, semantic links between identities, behaviors, and other contextual cues we will define to be *semantic associations* (or, for simplicity, just associations).

It should be noted that several sociological models define identity meanings in terms of similar kinds of semantic relationships. For example, Burke (1980) discusses the idea of role/counter-role pairs, i.e. those between "brother" and "sister", and generalizes this to the idea of identities and counter-identities, the notion that identities in general have one or more other identities on which we base their meaning.

However, unlike Burke (1980), cognitive psychologists have taken traits and associations and used them to construct predictive models for identity labeling. Most existing models are built within a parallel constraint satisfaction (PCS) framework. In PCS models, semantic links ("constraints") exist between nodes, which can be anything from identity labels to particular traits. Each node is ascribed a single attribute - a level of cognitive activation. Each node also starts with a base level of this activation value. Nodes can then be "excited" by external stimuli, at which point their base level of activation is increased by a set amount. Activation then flows through links in the system using a predetermined flow equation. Notably, links can either act as exciting links or inhibitory links. An exciting link between two schema means that activating one of the links will increase the activation of the other. An inhibitory link means that increasing activation in one schema will decrease the activation of another. The existence of these inhibitory links is the primary benefit of PCS models over pure spreading activation models (Collins and Loftus, 1975).

Several PCS models have been developed that show how semantic relationships inform the identity labeling process (Schröder et al., 2013; Freeman and Ambady, 2011; Kunda and Thagard, 1996). A representative model, to be focused on here, is that by Freeman and Ambady (2011). In the Freeman and Ambady model, nodes can be one of four types. Nodes at the cue level include visual and auditory features, such as an individual's face. At the category level, nodes indicating social categories (that is, a form of identity) exist. At the stereotype level, nodes exist that represent traits, such as annoying. Finally, a "higher-order" level includes nodes such as prejudice and motivations. Connections exist across nodes at different levels, and activation starting anywhere in the network is passed through the network until stability is reached, at which point an identity (category) is probabilistically selected based on its level of activation.

In the Freeman and Ambady model, then, the meaning space M of identities comes in the form of a network structure of semantic relationships. Loosely, the function ϕ spreads activation through this network and returns the activation level of each identity relative to the other potential options once activation levels within the network have reached a steady state. In Section 5, we formalize these intuitions mathematically to make predictions for the identity labeling problem.

An explicitly relational perspective of identity makes it much easier to model how overlaps and intersections between identities, both implicit and explicit, may manifest during the labeling process (Penner and Saperstein, 2013). Further, varying the strength and valence of these interrelationships can help to explain how, for example, biracial women can be more likely to identify as multiracial than biracial men (Davenport, 2016). The use of a cognitive, relational perspective therefore allows PCS models the ability to provide an appealing explanation for a variety of well-known phenomenon in the identity labeling process, in particular the existence of intersectional identities.

However, PCS models are context-specific; that is, they are designed to be hard-wired to how people label others in particular situations. This is because only one type of semantic relationship can be modeled in a given PCS model. A single model therefore cannot be used to demonstrate how a respondent might make multiple inferences based on the existence of a particular "link." For example, PCS models cannot represent simultaneously the fact that two distinct individuals standing together may be brother and sister (an "exciting link"), and that neither individual can be both a brother and a sister (an "inhibiting" link).

Additionally, modeling affective information in PCS models is difficult. PCS models rely on the implicit notion of an identity's affective meaning (e.g. implicitly, criminals are bad), but lack the expressiveness to represent this explicitly in the model. This is made most clear in considering how one might adapt behaviors into PCS models. Unlike cognitive associations we can represent between finite categorical variables, e.g. race and sex, identities can engage in a multitude of behaviors at any given moment. These behaviors, in turn, are more easily defined in terms of how they make us feel than by the cognitive associations they elicit. Further, as a situation evolves, the association of an identity to a behavior can change as we move through a rapid series of different behaviors. Consequently, PCS models with static link values and a static set of nodes, as the Freeman and Ambady model has, are infeasible for use

modeling how behaviors in a situation inform the identity labeling problem.

Schröder et al. (2013) address this issue by developing a PCS model of behavioral priming that draws directly on ACT. In their model, the PCS model's nodes are the Evaluation, Potency and Activity dimensions for different identity labels. Their model thus repurposes the PCS model within an affective framework, rather than explicitly combining cognitive and affective meanings as we seek to do.

Also noteworthy are the efforts of Ehret et al. (2014), who move beyond PCS models to more a generalizable, deep recurrent neural network structure for person construal. Ehret et al.'s model is an improvement over standard PCS models in that explicit links between concepts are replaced by latent representations of concepts and weights between latent representations that can be "learned" by training the model in a particular way. In principle, this could allow a model to represent both semantic and affective meanings in the future. Further, Ehret's model retains the benefit of existing PCS models, like the Freeman and Ambady model, of being able to represent semantic associations not only amongst identities, but also with higher-order cognitive concepts like motivations.

However, Ehret's model currently focuses only on the identity labeling problem in the first instant we attempt to label an individual we see visually. In contrast, we here seek a model that can address identity labeling that includes both labels of others in the situation and, more importantly, behaviors engaged in by identities. Additionally, we work here with theoretically driven latent representations of identities and behaviors, while Ehret et al. (2014) rely on generic latent structures that can learn potentially meaningful representations. Future work combining our efforts with those of Ehret et al. (2014), are, however, likely to be fruitful.

4 Survey Experiment

4.1 Overview

The prior section suggests three claims apply to existing predictive models of the identity labeling problem:

1. When salient information is predominantly given in the form of cognitive information—semantic associations and/or traits— a) PCS models should be predictive while b) (Bayes)ACT should struggle to make predictions

2. Where salient information come predominantly in the form of affective meaning derived from behaviors, a) (Bayes)ACT should be predictive while b) PCS models should struggle to make predictions

3. When substantial levels of both affective and semantic meanings are provided simultaneously, neither (Bayes)ACT nor PCS models should be predictive

In this section, we introduce a survey designed to address these claims. Specifically, we first use it to establish Claims 1a and 2a, that existing models of identity labeling can actually make accurate predictions about how people

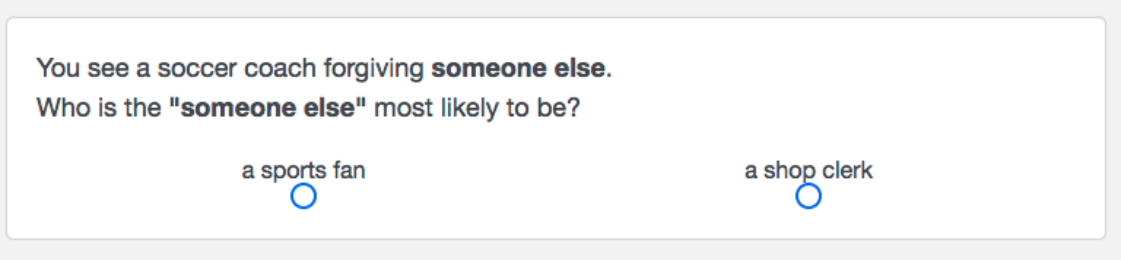

Figure 1: An example question from our survey

label individuals in hypothetical social situations. While prior work has suggested PCS models are capable of doing so in specific contexts, it is unclear how well they work across the variety of settings we test here. And while BayesACT has been used in a variety of fashions, no extensive test of it as an identity labeling model has yet been conducted.

The second purpose of our survey is to establish a motivation for a new identity labeling model that can incorporate affective, associative and trait-based meanings. To do so, we will consider failure cases in the spirit of Claims 1b and 2b. Combined, we thus expect to derive Claim 3, that when both affective and semantic meanings must be incorporated to make accurate predictions for the identity labeling problem, neither PCS models nor BayesACT models are sufficient. One solution to this issue is to combine PCS models with the capabilities of BayesACT. This is, practically speaking, what our LCSS model will do.

4.2 Survey Details

As shown in Figure 1, the core of our survey is a set of questions in which respondents are provided with a series of multiple choice questions. Each question gives a short statement describing an action that an individual ("someone else") takes or has taken towards them. We then ask the survey respondent (the labeler) to provide the best identity label for this individual (the labelee).

Such questions are a gross simplification of how identity labeling occurs in the real world, but these social-event based, binary-answer, minimally contextualized questions were selected for several reasons. The event-based nature of the questions allows us to inject affective meaning from behaviors into the questions in a theoretically principled way. Further, because we would like to compare our model to ACT's predictions, it is useful to work within the constraints of social events. We consider a binary response pattern in order to make evaluation easier - in the binary case, a model is simply "right" or "wrong", and cannot be "less wrong" (as is the case with more than two answers). Finally, by opting for minimal context, we can also minimize the extent to which respondents may be cued in ways we would not expect. In this way, we can control the salient information obtained by respondents and, consequently, assess how well the different identity labeling models we consider here are able to use the small set of signals we have made available.

In order to evaluate Claims 1, 2 and 3, we construct situations in which affect, associations and trait are differ-

Actor	Behavior	Answer 1	Answer 2	Affective Signal	Associative Signal
soccer coach / doctor	forgiving / assisting	soccer player / patient	shop clerk / cousin	Low	High (Role Pair)
soccer coach / doctor	hurting / punching	soccer player / patient	goon / trespasser	High	High (Role Pair)
soccer coach / doctor	forgiving / assisting	sports fan / paramedic	shop clerk / cousin	Low	Medium (Same Institution)
soccer coach / doctor	hurting / punching	sports fan / paramedic	goon / trespasser	High	Medium (Same Institution)
soccer coach / doctor	forgiving / assisting	goon / trespasser	sports fan / cousin	Low	Low (Different Institution)
soccer coach / doctor	hurting / punching	shop clerk / cousin	goon / trespasser	High	Low (Different Institution)

Table 1: Conditions for the Affective vs. Association questions

entially given as salient information to survey respondents. As noted above, traits and associations are generally not applicable to the same forms of the identity labeling problem (see the "brother and sister" example). Therefore, we consider two slightly different types of questions; one set which pairs different levels of affective and associative information (see Section 4.3), and the other that pairs varying levels of affective and trait cues (see Section 4.4). In this way, we more generally assess how models make predictions in the face of a combination of varying levels of affective and semantic bits of salient information.

4.3 Affective vs. Associative Survey Questions

Figure 1 provides a template for questions asked in this part of the survey, and Table 1 outlines the questions asked in more detail. In Figure 1, the Actor is soccer coach, the Behavior is forgiving, Answer 1 is sports fan, and Answer 2 is shop clerk. Each question is drawn from one of two different affective information conditions—"high" (salience) or "low" (salience)—and one of four different associative information conditions—"high" (association between the Actor and Answer 1), "medium", or "low". For each condition, we consider two different Actors, soccer coach and doctor, and construct questions around these actors based on the signal conditions.

The behavior the actor engages in is determined by the affective signal condition. At high affective signals, we select strong, negative actions (hurting and punching), while at low affective signal we select relatively innocuous ones (forgiving and assisting). Answer 2 is always selected to be the most affectively appropriate answer, as determined by Affect Control Theory. Answer 1 is instead determined by the associative signal Condition. At "high" levels of associative signal, answer 1 maps to a role pair for the actor (soccer players for soccer coaches, patients for doctors). At the medium level, answer 1 is selected to be an identity in the same institution. At the low level, we simply select a random identity. In all cases, we attempt to choose identities that do not strongly imply trait-based information, and thus control for trait.

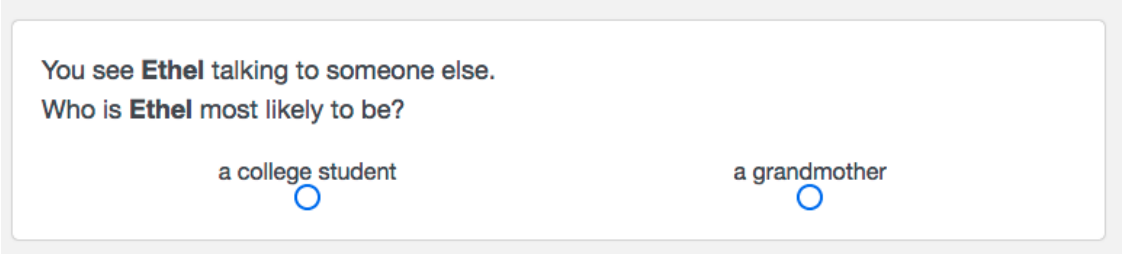

Figure 2: An example question from our survey for the Affect vs. Trait signal portion

Actor Name	Behavior	Object	Answer 1	Answer 2	Trait Signal	Affective Signal
Ethel / someone	attacking	enemy	bully	grandmother	Old,Female / None	High
Ethel / someone	talking to	person	college student	grandmother	Old,Female / None	Low
Harold / someone	attacking	enemy	bully	grandfather	Old,Male / None	High
Harold / someone	talking to	person	college student	grandfather	Old,Male / None	Low
Brittany / someone	attacking	enemy	villain	granddaughter	Young,Female / None	High
Brittany / someone	talking to	person	pharmacist	grandson	Young,Female / None	Low
Johnny / someone	attacking	enemy	villain	granddaughter	Young,Male / None	High
Johnny / someone	talking to	person	pharmacist	grandson	Young,Male / None	Low

Table 2: Conditions for the Affective vs. Trait questions

In addition to the twelve questions described in Table 1, we also include a set of 12 control questions. These control questions are identical to those described in Table 1, but instead of providing an actor identity instead simply use the term "someone" (i.e. "someone forgives someone else," who is that someone else?) These questions can be used to assess the affective model of ACT, absent any semantic meanings. Thus, these questions imply a fourth condition, the "none" condition, for associative meanings, that we test with our survey questions.

4.4 Affect vs. Trait Signals

Figure 2 provides a template for questions asked in this part of the survey, and Table 2 outlines the questions asked in more detail. Questions in this portion of the survey are distinct from the prior section in three respects. First, as opposed to asking the respondent to label "someone else", we ask them to label an individual with a name. We use the name information to encode trait information, and consider four different trait characteristics - "Ethel" is used to depict the traits "Old & Female", Harold for "Old & Male", "Brittany" for "Young & Female" and "Johnny" for "Young & Male".

Second, we here ask about the actor of a social event, as opposed to the object. Objects, and the behaviors taken towards them, are determined by the level of affective signal for the given condition, as shown in Table 2. Finally, note that we use two questions per condition, those shown in Table 2 as well as one with trait information removed. For example, for the "Old,Female" trait, high affective signal condition, we ask two questions - the one shown in Figure 2, as well as: "You see someone talking to someone else. Who is this someone most likely to be?" These questions serve as cases in which only affective signals are present.

In all cases, answer 2 is defined by matching the trait signal to an identity - for example, we match the identity grandson to the trait signal "Young & Male." Answer 1 maps to an identity that has a strong affective fit to the question and that also is distinct on the trait signal from Answer 2. For example, pharmacists may not be obviously male or female, but are likely to be implicitly characterized as being older than grandsons.

5 Modeling

In order to compare predictions across models, we must first formalize a common framework in which these predictions are comparable. We can then proceed to defining ACT, the PCS model of Freeman and Ambady and our new model, LCSS, fit into this framework. To begin, we state a slightly more formal definition of the identity labeling problem:

> Given a "labeler," x, and a "labelee," y, in a context with a particular set of active environmental cues r_t at time t, predict the set of identities $\mathbf{i_y}$ in the universe of all identities I that x will apply to y

While this general definition covers a broad range of phenomena, it is useful here to ground the problem specifically in the context of our survey data. Here, x is the survey respondent. In the Affective vs. Associative questions (Section 4.3), y is the "someone else". In our Affective vs. Trait questions (Section 4.4), y is the named individual. The universe of all identities I is limited to the two answer choices given for each question, and y is only ever labeled with a single identity i_y. Finally, in both sets of questions, the active environmental cues (i.e. the salient information given to the labeler) r_t are two-fold. First, for both kinds of questions, a social event provides salient affective information. Second, in the Affective vs. Association questions, the actor has varying degrees of semantic association with the answer choices. In the Affective vs. Trait conditions, the social event contains additional trait-based cues about y; a name and the traits the name implies.

Returning to the general case, we can derive from the problem statement above a mathematical form for how an identity labeling model uses M and ϕ to select a set of labels for x to apply to y. In doing so, it is reasonable to assume that an element of randomness is inherent in x's decision. Because of this, the result of an identity labeling model is best given as a probability distribution, where all identities have some (perhaps very small) chance of being applied.

Formally, the probability that x labels y with a randomly selected set of identities $\mathbf{i_a}$ given environmental cues r can be expressed as:

$$p(i_y = \mathbf{i_a} | \phi_x, r_{tx}, M_x) \tag{1}$$

In the context of our survey data, this expression says that every survey respondent x will have a different meaning space M, may observe or choose to attend to a distinct set of environmental cues r_t and may use this information in a distinct fashion from others (ϕ). In the present work, however, we will assume that $\phi_x = \phi, M_x = M, r_{tx} = r \, \forall \, t, x$. In other words, we assume that all survey respondents have the same meanings of identities and environmental cues, engage with the same set of cues and all use these meanings and cues in the same way to label y.

These assumptions are hopelessly incorrect. As one of many examples, the features of others that imply membership in a particular social category can vary depending on where we live and those that surround us **?**. In the present work, however, such simplifications are made for two reasons. First, our survey data is not large enough to estimate per-individual statistical models. Second, both Affect Control theorists and cognitive psychologists make similar assumptions in their existing identity labeling models, and the present work focuses on a comparison to and extension of these models.

Having defined a probability distribution of interest, we now can explicitly define how a generic identity labeling model might leverage a particular ϕ, M and r_t to define this probability distribution over identities. Assume that ϕ is a function with three parameters. First, ϕ takes some identity, i_a, to be "scored." Second, it takes in the meaning model of the theory, M. Finally ϕ takes the active environmental cues, r, from the current situation. The function $\phi(i_a, M, r)$ then assigns a score to the given identity for how likely it is to be applied in the current situation.

A scoring function on each element of a set can be turned into a probability distribution by way of a *discrete choice model* (DCM) McFadden (1980). We can therefore formally define $p(i_y = \mathbf{i_a} | \phi, r, M)$ via a discrete choice model (DCM) McFadden (1980), leading to the following:

$$p(i_y = i_a | M, r) = \frac{e^{\phi(i_a, M, r)}}{\sum_{j \in I} e^{\phi(i_j, M, r)}} \tag{2}$$

An identity labeling model can therefore be defined entirely the way it defines the scoring function $\phi(i_a, M, r)$. This is important, because it allows us to translate non-probabilistic predictive models, including ACT and the Freeman and Ambady models, into a consistent, probabilistic framework that can generalize to future identity labeling models.

Before continuing, one important note for the present work is that, in the special case of the identity labeling problem where the universe of possible labels, I, is a set of two options i_a and i_b (as in the survey data), we can further

reduce Equation 2. We show this below, where in the third step below we divide through by $e^{\phi(i_a,M,r)}$:

$$\begin{aligned} p(i_y = i_a|M,r) &= \frac{e^{\phi(i_a,M,r)}}{\sum_{j\in I} e^{\phi(i_j,M,r)}} \\ &= \frac{e^{\phi(i_a,M,r)}}{e^{\phi(i_a,M,r)} + e^{\phi(i_b,M,r)}} = \frac{1}{1+\frac{e^{\phi(i_b,M,r)}}{e^{\phi(i_a,M,r)}}} \\ &= \frac{1}{1+e^{\phi(i_b,M,r)-\phi(i_a,M,r)}} \end{aligned} \tag{3}$$

The final line in Equation 3 is a binary logistic model where the activation function is defined by the difference between ϕ for i_a and i_b. As we will show below, given a particular form of ϕ, this equation can be further simplified to the point where we can use standard logistic regression modeling to estimate parameters for identity labeling models.

We now provide a brief overview on $\phi(i_a, M, r)$ as defined by (Bayes)ACT, the PCS model of Freeman and Ambady and the new LCSS model.

5.1 (Bayes)ACT

5.1.1 Overview

In this section, we will largely focus on ACT, as BayesACT derives much of its mathematics from the original model. As noted above, ACT assumes a particular measurement system for the affective meanings of identities, and the behaviors these identities engage in. The meanings of these entities are defined as points in a three dimensional affective EPA space. This EPA space, and the positions of identities and behaviors within it, characterizes the meaning space, M, of ACT.

Environmental cues, r, that can be leveraged by ACT come in the form of social events, where an actor engages in a particular behavior towards an object. The identity labeling problem arises when we are not provided with the identity of the actor or the object, and must infer it from other information provided in the social event. For example, if we are told that "someone" taught a student, we can use the mathematics of ACT to predict the identity of "someone" given the affective meanings (in M) of "taught" and "student."

To do so, ACT relies on the idea of *affective deflection minimization*. ACT assumes that social events can change the meaning of identities and behaviors within a particular situation. Because humans generally expect these meanings to align with broader, static cultural norms, however, (Bayes)ACT theorizes that the best identity to label an individual with is the one that causes the smallest change, or deflection, in affective meanings. For example, assigning the identity "terrorist" to the "someone" in the "someone taught a student" example above would lead to high affective deflection, because we generally think of students and teaching as good and positive, respectively, while incorporating a terrorist into the situation may redefine our expectations of them in a negative light. In contrast, we generally expect to see

"instructors" teaching students, so this identity leads to low deflection.

In Section 5.1.2 below, we provide details of how ACT formalizes these intuitions mathematically. For those uninterested in the underlying mathematics, it should suffice that this intuitive deflection minimization process roughly defines a ϕ for the identity labeling problem. This is because it combines a meaning space M (distributions in EPA space of identities and behaviors) and some set of situational cues r (social event details) to produce a score for each identity, from which a probability distribution for labeling can be defined. In a sense, ACT, and BayesACT as well (as detailed below), can be thought of as giving the best possible identity labeling model assuming that M can be defined by EPA space and where affective deflection predicts how we label others.

5.1.2 Mathematics of (Bayes)ACT

ACT details how social events can also be used to infer labels for identities, and thus to solve the identity labeling problem. Here, we provide mathematical details on how this process works.

Social events have a pre-event transient meaning, f, for a given actor, behavior, object triplet. These pre-event meanings correspond to the fundamental, static, culturally-normed meanings in EPA space of the involved identities and behaviors. Upon a social event (or a series of events) occurring, these meanings are modified to produce a post-event transient meaning, τ. Both f and τ are vectors of length nine, one element each for the Evaluative, Potency and Activity affective dimensions for the $a$ctor, $b$ehavior and $o$bject, i.e. for f, we have:

$$f = \begin{bmatrix} a_e & a_p & a_a & b_e & b_p & b_a & o_e & o_p & o_a \end{bmatrix} \tag{4}$$

ACT specifies an equation that determines the values of τ as a function of f. This equation can be characterized by the form $\tau = \mathcal{Z}\, g(f)$, where the value of $g(f)$ is a kx1 vector of covariates and the matrix $\mathcal{Z}$ is a 9xk matrix specifying 9 different sets of regression coefficients, one for each element of τ. The actual values of $g(f)$ and $\mathcal{Z}$ are estimated via regression using survey data; the reader is referred to Morgan et al. (2015) for details on this process.

In the present work, we will assume that $g(f)$ and $\mathcal{Z}$ are given. We can then compute the post-event transient as follows, where $\mathcal{Z}_x$ represents row x of the coefficient matrix:

$$\tau = \begin{bmatrix} \mathcal{Z}_{a_e}^T \cdot g(f) & \mathcal{Z}_{a_p}^T \cdot g(f) & \dots & \mathcal{Z}_{o_a}^T \cdot g(f) \end{bmatrix}$$

Given f and τ, we can compute the *deflection* of a social event as the unnormalized Euclidean distance between the pre- and post-event transients, where the importance of each affective dimension can potentially be weighted by

some weight vector w. In practice, w_j is generally set to 1 for all elements of the fundamental:

$$deflection(f) = \sum_{j}^{9} w_j(f_j - \tau_j)^2 = \sum_{j}^{9} w_j(f_j - \mathcal{Z}_j \cdot g(f))^2 \tag{5}$$

Deflection gives an idea of how "expected" a social event is. An event where deflection is high indicates that the event significantly changes the affective meanings of the actor, behavior and object. Because ACT expects these meanings to be relatively consistent over time, this implies that the described social event is unlikely. Similarly, a low deflection signifies an event that "makes sense."

The above equation can therefore be used as a ϕ function for the identity labeling problem in the case where an actor (object) with an unknown identity enacts (receives) a behavior on (from) an object (actor) with a known identity. In the language of the identity labeling problem, where we require a function $\phi(i_a, M, r)$, we can say that i_a is a potential label (to be scored) for the unlabeled actor (or object), M is the culturally defined EPA space given by ACT, and r is the behavior the actor (object) is engaging in (receiving) as well as the identity of the object (actor) that i_a is acting (being acted) upon. The only distinction to be made is that, because events with higher deflection suggest more *un*expected events, and ϕ should be higher with identities that better fit the situation, we negate the deflection score in our construction of ϕ.

Let us assume we wish to identify an unlabeled actor. Then for each potential identity in I, i_a, that we wish to score, we have environmental conditions r composed of a behavior, b and object, o, and meanings derived from EPA space. This gives the following, where, e.g., M_{x_e} represents the Evaluative dimension of entity x and $\circ$ concatenates two vectors:

$$\begin{aligned} \phi(i_a, M, r) &= -deflection(i_a \circ r) \\ \text{where} \quad i_a &= \begin{bmatrix} M_{i_{a_e}} & M_{i_{a_p}} & M_{i_{a_a}} \end{bmatrix} \\ \text{and} \quad r &= \begin{bmatrix} M_{b_e} & M_{b_p} & M_{b_a} & M_{o_e} & M_{o_p} & M_{o_a} \end{bmatrix} \end{aligned} \tag{6}$$

By computing deflection across all identities, i_a, in I and then substituting in to the discrete choice model in Equation 2, we can construct the desired probability distribution for the identity labeling model. This is exactly the procedure we employ here, using regression equations and EPA values defined by Heise (2007) and provided in the code and data release for this article. The only exception to this are the EPA values of the Actor names in Table 2, for which no EPA values exist. For these, we provide our own survey respondents with traditional questions used to estimate EPA values, and use the mean values of these responses as EPA values for these concepts.

This completes the definition of ACT as an identity labeling model; note, however, that we have not addressed

its successor, BayesACT. In the Appendix, we describe how the identity labeling framework we describe here closely parallels the mathematics of BayesACT, and why we chose to use the simpler, more straightforward ACT model here.

5.2 Freeman and Ambady's Semantic Model

5.2.1 Overview

The meaning space M of the Freeman and Ambady model consists of a cognitive network of associations between concepts. Environmental cues are given to the model by way of the activation of particular concepts in a given social situation. In the original article, activation of a concept comes in two forms. First, the labelee can emit some observable information (e.g. skin color, or "being annoying") that directly triggers activation of a concept. Once concept activation occurs, activation can then be spread to a particular concept through the cognitive network in a manner defined by a set of update equations.

Operationalization of Freeman and Ambady's model therefore would formally require that we specify concepts of interest and how they are activated by survey questions (r), how these concepts are connected in a cognitive network (M) and how activation spreads through that network (ϕ). Unfortunately, construction of M is difficult, often relying on researcher intuitions (Freeman and Ambady, 2011; Kunda and Thagard, 1996; Ehret et al., 2014).

Such a construction might be feasible for our survey experiment, given that our survey questions are relatively uncontextualized and thus provide a fairly small space of potential constructions for M. In our Association vs. Affect survey questions described in Section 4.3, the potential network structures only can involve the actor identities ("soccer coach" and "doctor") and any identities given as potential choices for object identities. In our Trait vs Affective questions (Section 4.4), semantic relationships are restricted to the connection between the traits old, young, male and female, and each identity.

Consequently, we might approximate M and $\phi(i_a, M, r)$ in these simple networks by trying to define the differing strengths of semantic relationships between the identity already labeled and the other identities in the survey question. More directly, however, we can simply approximate the *output of* ϕ, thus assuming but not explicitly needing to describe some reasonable semantic network structure (M) and update equations (ϕ). For example, without needing to construct any specific M, we can make reasonable assumptions about the final activation state of the identities "soccer player" and "goon" given that the concept "soccer coach" has been activated. Because these final activation states are all that is needed to make determinations on the identity labeling problem, we need not worry about exactly the form that M would take.

In fact, we can not only make assumptions about these values, we can actually learn the most likely weight of an unknown meaning space M given the assumptions of the Freeman and Ambady model and *the survey data itself*. More specifically, we can estimate the *best possible* semantic model that could be constructed given the assumptions of PCS

models, and then test the ability of these estimates to make predictions. In a broad sense, this model provides the best possible model under the assumption that semantic associations and/or traits are the only thing necessary to model to explain how we label others under the assumptions of a PCS model. Below, we explain in more mathematical detail how we formalize this mathematically.

5.2.2 Mathematics of Freeman and Ambady's model

As noted above, the Freeman and Ambady model assumes that the meaning space M is described by some semantic network N. The network N consists of a set of concepts, or nodes N, and a set of edges between these concepts E. Some of the nodes are the identities in I, others might be relevant to specific traits (e.g. brown hair), settings (e.g. courtroom), motivations, etc. A positive edge is placed between concepts that "stimulate" activation of the other - for example, the setting "school" stimulates the activation of the identity "teacher". Similarly, a negative edge is placed between concepts that "inhibit" activation of the other - for example, the trait "young" might inhibit activation of the identity "grandfather".

Environmental cues, r, are given in the form of one or more concepts that are initially activated in a given situation. This activation then spreads through the network via a series of iteratively applied update equations across links. When two nodes share a positive edge, activation in one node increases the level of activation in the other. When two nodes share a negative link, activation in one node decreases activation in the other.

We refer the reader to the appendix of Freeman and Ambady (2011) for the exact specifications of these update equations. Here, we note only that after some time, the level of activation for each node settles to a stable value representing how cognitively activated that node is, given the original stimuli r flowing through the network N. These update equations, iteratively applied, thus specify the function ϕ. The Freeman and Ambady (2011) model then assumes that the likelihood of a particular identity i_a being used to label an individual is proportional to the level of activation of i_a versus all other identities in I.

To see how this can apply to our survey data, let us focus on the set of Affective vs. Association questions. Here, the environmental cue consists of the already-labeled individual in the social event, i_q, and a behavior b. As noted above, behaviors selected for the present work function to provide affective and not semantic information for the identity labeling problem. Thus, we assume all behaviors we study here are isolates in the network N, and therefore cannot spread any activation. Consequently, we can focus only on the impact of i_q here.

Let us now assume that we have, in some way, constructed a ϕ and a network N - perhaps we use the original form of ϕ in Freeman and Ambady (2011) and N consists of a network linking i_a and i_b, the two potential choices for our survey respondent, and i_q all together with positive links (potentially of varying weights). We then input i_q as our environmental cue r and generate final activation rates for i_a and i_b.

Let us call these final rates of activation β_a for i_a and β_b for i_b. Given this, and the result in Equation 3, we have that:

$$p(i_y = i_a|M, r) = p(i_y = i_a|N, i_q) = \frac{1}{1 + e^{\phi(i_b,N,i_q)-\phi(i_a,N,i_q)}} = \frac{1}{1 + e^{\beta_b-\beta_a}} \tag{7}$$

From this, we can see that if we can determine sensible values for β_a and β_b, we need not worry about the exact specification of ϕ and N for our survey data. This is useful for two reasons.

First, as noted above, the meaning space (semantic network) N for Freeman and Ambady (2011) and similar models is generally constructed by hand. Consequently, the values of β_a and β_b are characterized fully by manually-defined values anyway, so we can always construct a ϕ and N to generate any values of β_a and β_b that we might deem reasonable. Second, and more importantly, is that by assuming we are only interested in these final values, we can use our data to train the best possible values for β_a and β_b with respect to their consistency with the survey data. Thus, we can develop the best possible predictions under the assumptions that semantic associations, and not affective meanings, are what drive identity labeling decisions. Although we do not focus on this in the present work, it is worth noting that as a byproduct, we can also provide estimates of the activation level of each identity in our survey data given an i_q.

To do so, let us first assume that β is a vector of length K, where K is the number of identities (4) in each of the two scenarios presented in Table 4.4. These scenarios are distinguished by particular i_q - in the first, i_q is "soccer coach", in the second, it is "doctor". In the first, then, we estimate β for the identities "soccer player", "sports fan", "shop clerk" and "goon", in the second, for "patient", "paramedic", "cousin", "trespasser".

Let us now focus only on the first scenario, as we will estimate β for these two scenarios separately. We now define two binary vectors a and b, which tell us what identities are provided as i_a and i_b for a particular question. Because there are four total identities in each scenario considered, a and b are of length four. Let us now assign an arbitrary "position" for each identity across any question answer in each scenario. So, for example, "soccer player" is assigned to position 1, "shop clerk to position 3, etc.

When we ask "A soccer coach is assisting someone, is that someone a soccer player or a shop clerk" (first row, first scenario in Table 1), then $a = \left[1000\right]$ and $b = \left[0100\right]$, and so on for all other questions provided. We can now define the probability that a given survey respondent selects i_a for all survey questions for this "soccer coach" scenario:

$$\phi(i_y = i_a) = \frac{1}{1 + 1 + e^{\beta^T b-\beta^T a}} = \frac{1}{1 + 1 + e^{\beta^T (b-a)}} \tag{8}$$

Carried across all survey respondents and all questions for this scenario, this is exactly a logistic regression model where the data is $b - a$ and we estimate the final activation values for each identity answer choice given that i_q is

the stimulus. Therefore, we can estimate the *most likely final activation score* under the Freeman and Ambady model *directly from the data*. We can run an analogous procedure for the second scenario in the Affective vs Associative set of questions.

For the Affective vs. Trait-based set of questions, we can use a similar approach. Again, we can assume that there is some N and ϕ that details how the traits implied by the names that we provide respondents (e.g. "Ethel") match each identity given as an answer (intuitively, "grandmother" but not "bully"). We could then estimate a single β, which is the difference in activation dependent upon whether or not a trait is activated by a given answer choice.

This strategy for estimation is, however, contingent on the expectation that the names provided would match the gendered and age-related demographics we expected them to align with. We included a manipulation check for the connection of these names to the associated ages by asking respondents how old they expected someone named "Ethel", "Harold", "Brittany" and "Johnny" to be. On average, Ethel was expected to be around 73 years old; Harold around 60, Brittany around 29 and Johnny around 30.

This suggests that our age manipulation worked much better for Ethel, and to a certain extend Harold, than it did for Johnny and Brittany. From a PCS model perspective, the name Ethel most strongly activates the trait "Old" (as, presumably, does "grandmother"), while in contrast, "Johnny" does not as strongly activate the youth trait we would expect to be associated with the identity "grandson". Because of this, we decided to provide the PCS model with additional information on the strength of the trait association based on our intuitions derived from the manipulation check. Specifically, we "told" the model that the activation of the "Old and Female" trait by Ethel was three times as strong as the corresponding activations of Brittany or Johnny to the "Young and Female" or "Young and Male" traits, and that the activation of the "Old and Male" trait by Harold was twice as strong.

5.3 Latent Cognitive Social Spaces

5.3.1 Overview

If, in the context of the identity labeling problem, only affective meanings of social events were relevant, (Bayes)ACT would provide an ideal (or nearly ideal) predictive model. It explains how affective meanings defined by a social situation allow us to identify a set of likely identities for an unlabeled individual. Similarly, if identity labeling decisions were made only on the basis of cognitive meanings, or semantic relationships including associations and traits, Freeman and Ambady's model would be likely to provide the most accurate predictions.

If, however, it were to be some combination of affective and semantic meanings that determined how we label others, neither model would be sufficient. (Bayes)ACT currently contains no semantic constraints in its predictive model. Similarly, while affect could in theory be included into PCS models, the mathematical nor conceptual underpinnings of how to do so currently exist. More specifically, PCS models are not currently capable of explaining how affect

generated from general social behaviors shapes identity labeling. Given that ACT and PCS models, respectively, have shown that both affective and semantic signals are important in the identity labeling problem, a logical step forward is to construct a single model that combines affective, associative and trait-based environmental cues.

We introduce Latent Cognitive Social Spaces to fill this void. Theoretically, LCSS assumes the same control process model as BayesACT on the meanings of identities and behaviors (Robinson, 2007). In other words, LCSS assumes that there are fundamental meanings of identities and behaviors that we work to maintain in social situations. However, instead of maintenance of only affective meanings, LCSS proposes that there are also culturally shared meanings of trait-based and association-based meanings of identities (only) that people attempt to retain when interacting socially or when observing social interactions. It is the combination of these meanings and their predictive power that provide a full explanation for the identity labeling problem. In the mathematical language we have developed here, LCSS thus leverages a very similar ϕ as that posed by (Bayes)ACT, but diverges in its definition of M. In the Discussion section, we touch in more detail the theoretical implications of this decision. Here, we focus on how we can develop a predictive model using these assumptions.

Incorporating trait-based and association-based meanings into the deflection framework of ACT is difficult for two reasons. First, existing models of identity labeling that incorporate traits and associations almost universally focus on network-based semantic models. In contrast, the deflection equation centers on the notion of comparing positions in a latent geometric space. In order to address this, LCSS proposes that semantic associations and traits, while perhaps best modeled as a network, can be approximated by an adequate geometric space. This idea is not new; network scholars have for years either implicitly (by visualization) or explicitly (via latent space statistical models Hoff et al., 2002) represented networks spatially. Thus, while our model adopts a solely latent space approach, it is conceptually similar to simply "smushing together" network-based PCS and latent-space based ACT models of meaning.

Second, while ACT scholars have developed decades worth of survey data that a) suggest affective meanings can be captured in three dimensions and that b) provide dictionaries of affective meanings for thousands of identities and behaviors in EPA space, it is less clear exactly how many trait-based and association-based "dimensions" we need, and further how identities exist within that space. For example, are status characteristics the only trait-based dimensions needed? And if so, where does the identity "soccer coach" fall in this space? Similar questions could be asked of associative dimensions - Heise and MacKinnon's (2010) institutional model seems relevant, but do we need the full set of these dimensions in order to faithfully model the identity labeling process?

The present work leaves these questions largely unanswered. Instead, we simply instantiate a sensible formulation of the model for the simplistic data and form of the identity labeling problem we study here. We then show that by combining cognitive and affective meanings into a single model, we can improve our ability to predict labeling decisions by survey respondents.

To make these predictions, as with our approach to PCS models, leverage the simplicity of our survey design and

the mathematics of our theory to determine the best predictive model, as estimated via some subset of our data, under the assumptions of LCSS. If this model improves over predictions of ACT and the Freeman and Ambady PCS model, we can have confidence that the assumptions of LCSS are more correct, at least for this particular problem.

Below, we provide mathematical details for LCSS. Essentially, the model we propose attempts to balance affective, trait-based and association-based meanings to make the best prediction for the identity labeling problem. For example, when there is little affective meaning given by a situation, the model relies more on associations and traits; in contrast, when affective information dominates, associations and traits are essentially ignored by the model. How strongly we weight these different dimensions is determined by a statistical model that incorporates existing data from ACT, rough approximations of the importance of various forms of semantic associations and traits, and weight estimators determined from a portion of our survey data.

5.3.2 Mathematics of LCSS

LCSS leverages the same concept of deflection defined by ACT (Equation 5). Like ACT, we can use this idea of deflection to define a ϕ for the identity labeling problem. The primary difference between the ACT formulation of the identity labeling model and the LCSS formulation, then, is in M, the set of requisite meanings. Specifically, LCSS assumes that in addition to affective meanings, identities have trait-based meanings and association-based meanings as well. A subtly that bears repeating here is that we assume *only* identities, and *not* behaviors, have positions in trait and association latent spaces. This assumption is not necessary for mathematical purposes but is theoretically appealing in that we believe few behaviors can be readily semantically linked to traits or associations beyond what affective meanings may imply.

Affective, trait-based and association-based meanings are characterized by a position in a multi-dimensional latent space. The first three dimensions of this space are affective, and define EPA space. All identities and behaviors are given a position in this affective space. The second $|T|$ dimensions of this space define meaningful traits upon which culturally shared definitions exist. Trait dimensions are defined by two opposing "poles" along which identities are culturally aligned for a particular trait. For example, the "age" dimension would have two poles, "young" and "old", and the sex trait would have poles for "male" and "female". Each identity holds a position in this trait space; for example, "soccer coach" might be closer to the "old" end of the age dimension, and near the middle of the sex trait.

Finally, the last $|K|$ dimensions of the meaning space for LCSS define dimensions along which we can reconstruct a network of semantic associations in a latent space. In general, statistical techniques exists to rigorously define the number of latent dimensions needed to faithfully capture properties of a given network (Hoff et al., 2002). In the present work, it suffices to think of these associative dimensions as characterizing the strength of association of each identity to a set of latent social institutions (Heise and MacKinnon, 2010). For example, one dimension might model association

with the educational institution, another with the judicial system, and so on. Collectively, this set of $3 + |T| + |K|$ latent dimensions, and a set of specified positions of behaviors and identities within them, defines the meaning space M for LCSS.

To incorporate these new meanings into a deflection model that we can use to "score" a potential identity for the labelee, we must augment the existing deflection equation from ACT. First, the fundamental and transient vectors of ACT must be expanded. To do so, we will require some additional notation. As in Equation 4, let a, b and o stand for the *a*ctor, *b*ehavior and *o*bject in a social event, and let, e.g., a_e stand for the evaluative dimension of the actor. Additionally, let f and τ hold the same meanings as above. Finally, as implied above, let $t_0, t_1, ..., t_T$ be a set of T traits for which fundamental meanings exist, and $k_0, k_1, ..., k_K$ be a set of K associative dimensions for which fundamental meanings exist. Given this notation, we now define f^* to represent the fundamental vector for LCSS:[1]

$$
\begin{aligned}
f^* &= f \circ f_t \circ f_k \\
&\text{where} \quad f_t = \begin{bmatrix} a_{t_0} & a_{t_1} \ldots & a_{t_T} & o_{t_0} & o_{t_1} & \ldots & o_{t_T} \end{bmatrix} \\
&\text{and} \quad f_k = \begin{bmatrix} a_{k_0} & a_{k_1} \ldots & a_{k_K} & o_{k_0} & o_{k_1} & \ldots & o_{k_K} \end{bmatrix}
\end{aligned}
$$

Having described a new fundamental, we now must characterize the way in which fundamental meanings are changed by an observed social interaction. Mathematically, we will need to define the quantities $\mathcal{Z}*$ and $g^*(f^*)$, the matrix of regression parameters and coefficients, respectively, that are needed to determine transient meanings from fundamental meanings. For reasons explained below, we also will define a new weight vector, w^*, that is of length $|f^*|$. Using these variables, Equation 9 gives the deflection equation for LCSS:

$$
deflection_{lcss}(f^*) = \sum_{j}^{|f^*|} w_j^*(f_j^* - \tau_j^*)^2 = \sum_{j}^{|f^*|} w_j^*(f_j^* - \mathcal{Z}^{*T}_j g^*(f^*))^2 \tag{9}
$$

With no data to base our estimates on, it is impossible in the present work to empirically characterize the new parameters $\mathcal{Z}^*$, $g^*(f^*)$ and w^*. However, as we will show, reasonable assumptions above the form of $\mathcal{Z}^*$ will lead us to a tractable model we can use to make predictions on our survey data. Specifically, we will assume that affective, trait-based and association-based meanings are *independent* - that is, for example, the affective meanings of a before an event are unrelated to the transient association or trait-based meanings of a, b or o. Regardless of the form of $g^*(f^*)$, this implies that $\mathcal{Z}^*$ is structured as follows, where $\mathbf{0}_l$ specifies a zero vector of length l and, as above $\mathbf{Z_j}$ represents

[1]For ease of explanation, let us here assume these are points in a space, as in ACT, rather than distributions, as in BayesACT. However, we note that similar Bayesian extensions to LCSS could, in principle, be developed

Experimental Relationship	**Value of** $deflection_K(f_k)$
High (Role Pair)	0
Medium (Same Institution)	1
Low (Different Institution)	3
No Semantic Info (i_q is "someone")	0

Table 3: Assumed deflection values for Association-based deflection

the jth row of the coefficient Z matrix where we assume that we have values for trait and association dimensions:

$$\mathcal{Z}^* = \begin{bmatrix} \mathbf{Z_{a_e}} & \mathbf{0}_{|T|} & \mathbf{0}_{|K|} \\ \mathbf{Z_{a_p}} & \mathbf{0}_{|T|} & \mathbf{0}_{|K|} \\ \cdots & \cdots & \cdots \\ \mathbf{Z_{o_a}} & \mathbf{0}_{|T|} & \mathbf{0}_{|K|} \\ \mathbf{0}_{|f|} & \mathbf{Z_{a_{t_0}}} & \mathbf{0}_{|K|} \\ \cdots & \cdots & \cdots \\ \mathbf{0}_{|f|} & \mathbf{Z_{a_{t_T}}} & \mathbf{0}_{|K|} \\ \mathbf{0}_{|f|} & \mathbf{0}_{|T|} & \mathbf{Z_{a_{k_0}}} \\ \cdots & \cdots & \cdots \\ \mathbf{0}_{|f|} & \mathbf{0}_{|T|} & \mathbf{Z_{a_{k_K}}} \end{bmatrix} \tag{10}$$

This form of $\mathcal{Z}^*$ allows us to split the deflection equation in Equation 9 into three separate and independent deflection computations, as shown in Equation 11 below, where $deflection_T$ and $deflection_A$ define the subset of the deflection calculation relevant to traits and associative dimensions, respectively:

$$deflection_{lcss}(f^*) = \frac{w_f}{|f|} deflection(f) + \frac{w_{f_t}}{|f_t|} deflection_T(f_t) + \frac{w_{f_k}}{|f_k|} deflection_K(f_k) \tag{11}$$

Equation 11 shows that while we can express the LCSS model as a single, intuitive theoretical construct, we can at the same time, with a reasonable assumption, decouple the three different components of the model for use in an identity labeling context. This is convenient for the present work, as if we are able to define reasonable assumed values for $deflection_T(f_t)$ and $deflection_K(f_k)$ for the simplistic survey questions asked, we can then leverage a similar logistic regression framework as in the PCS models above to estimate the relative weights of trait-based, association-based and affective deflection in computing $deflection_{lcss}$.

Let us first explain how we characterize associative deflection, $deflection_K(f_k)$, for the various survey questions. For Trait vs. Affect questions, we provide no associative information. Consequently, we assume $deflection_K(f_k) = 0$ for all of these questions, or equivalently, that $deflection_K(f_k)$ is the same for both answer choices. For Association vs. Affect questions, we vary the deflection according to the different conditions for each answer choice as given

in Table 3. As we performed no manipulation checks on semantic associations, these values were set once, prior to data analysis. While this almost certainly diminishes the power of our predictive model, it also allows us to avoid potential "researcher degrees of freedom" in our analysis. The quantities in Table 3 state that high association leads to an "expected" interaction, and thus low deflection. In contrast, when no semantic information is provided, we cannot expect any association-based deflection to occur. Consequently, deflection is 0, and thus all deflection for these events draws only from affective meanings.

Now, let us turn to a definition of trait deflection, $deflection_T(f_t)$ for the various survey questions. For the Association vs. Affect questions (Section 4.3), no trait-based information is provided. Therefore, for all questions of this kind, we assume $deflection_T(f_t)$ is the same for both answer choices. For the Trait vs. Affect questions, we will assume that as trait "match" increases, trait-based deflection decreases. Where no trait information is provided (when the actor is "someone"), we will again assume that $deflection_T(f_t)$ is the same for both answer choices. For questions where trait information is provided, we will assume that trait deflection of the answer not aligning to the trait (i.e. "bully" in the question where the actor is Ethel) is K times greater than trait deflection for the trait-conforming answer ("grandmother"). We set K using the quantities for each name in the PCS model description section, motivated by our manipulation check (e.g. "grandmother" causes 3 times less trait deflection than "bully" for the actor Ethel).

Having specified values for $deflection_K(f_k)$ and $deflection_K(f_k) = 0$, we can complete the steps necessary to constructing a ϕ for LCSS by using the negation of the deflection equation in Equation 11. This is analogous to the approach we take for ACT in Equation 6, for brevity we do not repeat it here. Therefore, we have the following for LCSS for our survey questions, where we use d as a shorthand for $deflection$, e.g. $deflection_{lcss} = d_{lcss}$, and $f^*_{i_a}$ is the fundamental vector for a given question with i_a's values

$$\begin{aligned}
\phi(i_y = i_a) &= \frac{1}{1 + \exp\left(\phi(i_b, M, r) - \phi(i_a, M, r)\right)} = \frac{1}{1 + \exp\left(d_{lcss}(f^*_{i_a}) - d_{lcss}(f^*_{i_a})\right)} \\
&= \frac{1}{1 + \exp\left(\frac{w_f}{|f|} d(f_{i_a}) + \frac{w_{f_t}}{|f_t|} d_T(f_{t,i_a}) + \frac{w_{f_k}}{|f_k|} d_K(f_{k,i_a}) - \frac{w_f}{|f|} d(f_{i_b}) - \frac{w_{f_t}}{|f_t|} d_T(f_{t,i_b}) - \frac{w_{f_k}}{|f_k|} d_K(f_{k,i_b})\right)} \\
&= \frac{1}{1 + \exp\left(\frac{w_f}{|f|}(d(f_{i_a}) - d(f_{i_b})) + \frac{w_{f_t}}{|f_t|}(d_T(f_{t,i_a}) - d_T(f_{t,i_b})) + \frac{w_{f_k}}{|f_k|}(d_K(f_{k,i_a}) - d_K(f_{k,i_b}))\right)}
\end{aligned} \tag{12}$$

The only unknown values in Equation 12 are the weights on each of the separate forms of deflection. As in Equation 8, we can therefore estimate these weights from survey data, giving an indication of how important each of these factors are in predicting how individuals will label others. In theory, we can use these weights not only for prediction, but to give theoretical grounding to the importance of affect, trait-based and association-based signals in how people label others. In practice, however, we can only do so to the extent that the values we set for associative and trait-based deflections are accurate. Given that we provide only rough approximations for these quantities, we do not

explore this direction in the present work.

5.4 Model Evaluation

We carried out the survey described above to evaluate model performance with 80 participants on Amazon's Mechanical Turk.[2]

The ACT, Freeman and Ambady and LCSS models we have constructed all generate a probability for each potential identity to be selected. In order to assess the quality of these probabilistic predictions, we compute a single score, which we will call $prop\_answer1$, for each survey question. We compute $prop\_answer1$ by calculating the percentage of the 80 survey respondents that selected Answer 1 for each question. So, for the question given in Figure 1, $prop\_answer1$ is the percentage of respondents who selected the answer "a sports fan." As detailed below, we can then use the values of $prop\_answer1$ for each survey question in various ways to assess the quality of model predictions.

Finally, we note that because we perform only internal comparisons and do not focus heavily on absolute error measures, the PCSM and LCSS models are estimated and evaluated on the same dataset. Further, ACT uses no parameters estimated from the survey data, putting it at a slight disadvantage. Our attempts to address this, by using survey data to estimate weights for ACT in a similar fashion to LCSS, did not lead to significant improvements; these results are thus omitted for simplicity.

6 Results

In this section, we first use the results of the survey to assess how well the basic claims we made in Section 4.1 about existing models hold in the survey data we collect. We then turn to a comparison of the predictions of ACT and the Freeman and Ambady model to the predictions of our new LCSS model. Although it is only am example of a broader class of PCS models, we will refer to our instantiation of the Freeman and Ambady model with the shorthand "PCSM" in the results section for simplicity.

6.1 Assessing basic claims about existing models

In Section 4.1, we proposed several claims that we would expect to hold empirically given what we know about the meaning space (M) and "fit function" (ϕ) put forth by ACT & PCSMs. Here, we show that our empirical data is consistent with most, but not all, of these claims.

Claim 1 stated that a) PCSM should perform well when salient cognitive signals are given and salient affective signals are minimized, and b) ACT is likely to perform poorly in such situations. Because absolute performance is

[2]Subjects were restricted to those living in the United States that had completed 500 or more tasks (or "HITS") previously on Mechanical Turk with an acceptance rate of 95% or higher.

hard to assess (i.e. determining what is "good" performance is difficult), we evaluate only the relative nature of this claim - that PCSMs should perform much better than ACT under these conditions.

In order to test this assumption, we calculated the *mean absolute error (MAE)* of each model's predictions over the eight survey questions in the low affect condition across both surveys and that had either some form of trait information or medium to high levels of semantic association. The MAE represents an average, across all questions, of the difference between what percent of people a model thought would choose Answer 1 and the number of people who actually selected Answer 1.[3] The MAE of ACT over the eight low affect and medium/high cognitive signal conditions is 43.1%, more than two and a half times that of PCSMs (16.1%). This difference is statistically significant ($p < .01$).

Claim 2 stated that a) ACT should perform well when no salient cognitive information is provided and that b) PCSMs should perform poorly in this case. Again, empirical results confirm our intuitions. Here, we consider MAE for the twenty questions where no trait or semantic association information was provided. The MAE for ACT on these twenty questions was only 12.4%, nearly two and a half times as better than the MAE of the PCSM (29.4%).

Finally, Claim 3 argued that when substantial levels of both affective and semantic meanings are provided simultaneously, both ACT and PCSM would struggle. Here, empirical results support our claim relatively but not absolutely. We assess Claim 3 by looking at the MAE of ACT and PCSM for the eight questions where affective information was high and where trait information was provided or high to medium semantic association information was given. The MAE of ACT on these questions was 19.6%, for PCSM, 16.1%. These performances are not practically different from those we considered for Claim 1a) (for PCSM) and 2a) (for ACT) above. Empirical evidence therefore suggests that ACT and PCSMs perform worse but not overwhelmingly so in situations where both affective and cognitive signals are high.

That these models perform reasonably well in such situations does not mean, however, that they agree. To the contrary, Figure 3 shows that in all eight questions where strong affective and cognitive information is given to survey respondents, ACT and PCSM make opposite predictions. ACT constantly under-predicts how often respondents choose Answer 1, favoring Answer 2 which, by experimental design, was the better affective fit. In contrast, because only Answer 1 provides salient cognitive information, PCSM tends to overpredict the likelihood of respondents selecting that answer. We now turn to an investigation of the extent to which our new model, LCSS, can improve upon these predictions.

6.2 Model Comparison

Figure 4 presents predictions for the three different models considered suggests that LCSS provides the best fit to the survey data. Visually, Figure 4 suggests that LCSS performs better than the other two models because its estimates

[3]Formally, MAE is computed for a model k and a set of questions Q as $MAE(k) = \frac{1}{|Q|} \sum_{q \in Q} |p_k(q) - prop\_answer1_q|$, where $p_k(q)$ is the model's prediction for survey question q.

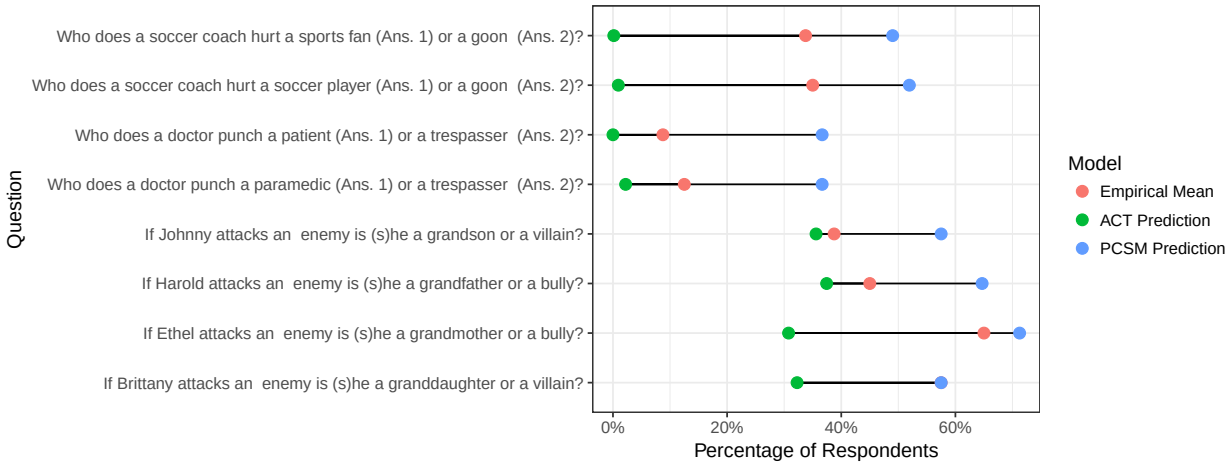

Figure 3: On the y-axis, the text of the question provided to survey respondents. The x-axis gives the percentage of respondents either estimated (by a model) or who actually chose the first answer ($prop\_answer1$). The different color plots represent results for the two models (Green for ACT, Blue for PCSM) and the empirical prediction (Red). Note that for the final question in the plot, the blue and red dots overlap (and so the red dot is hidden.

"pull" outliers in the ACT and PCSM models closer to the diagonal black line. For example, as suggested by analysis of Claim 1 above, outliers in Figure 4a) occur when affective information is low but semantic information is not. Because LCSS weights both affective and cognitive information, however, it is able "know" that where semantic information is prevalent, it is leveraged by respondents. Similarly, where no trait information is provided, PCSM models assume Answer 1 and Answer 2 are equally likely, in contrast, LCSS captures affective information to capture important differences between the two identities provided to respondents. Although LCSS has several more parameters than the PCSM model, it is notable that LCSS improves performance over PCSM models while estimating fewer (Affect vs. Association) or only one additional (Affect vs. Trait) parameter(s) from the data itself.

What Figure 4 does not suggest is that LCSS provides perfect predictions; a host of answers provided are significantly different from survey responses. Additionally, it appears that improvement from LCSS seems to largely derive from cases where ACT leans heavily on limited affective information (the low affect conditions) or where the PCSM has limited or no cognitive information to draw on. These two cases suggest that LCSS helps to ameliorate issues implied by Claims 1 and 2 above. Where both signals exist, however, LCSS does not obviously improve upon predictions of the two models. Why LCSS is unable to better blend cognitive and affective signals in these cases is likely due in part to the small sample size we use here, but suggests an existing limitation of the model.

Still, improvements made by LCSS in combining cognitive and affective meanings are clear. In addition to the visual evidence in Figure 4, we also calculate the MAE over all questions for PCSM, ACT and LCSS. LCSS, on average, is off on its predictions by 11.5%, with a 95% bootstrapped confidence interval of [8.8%-14.2%]. Both ACT (19.7% [14.8%-25.3%]) and PCSM (25.0% [20.4%-30.1%) perform significantly worse. In absolute terms, LCSS

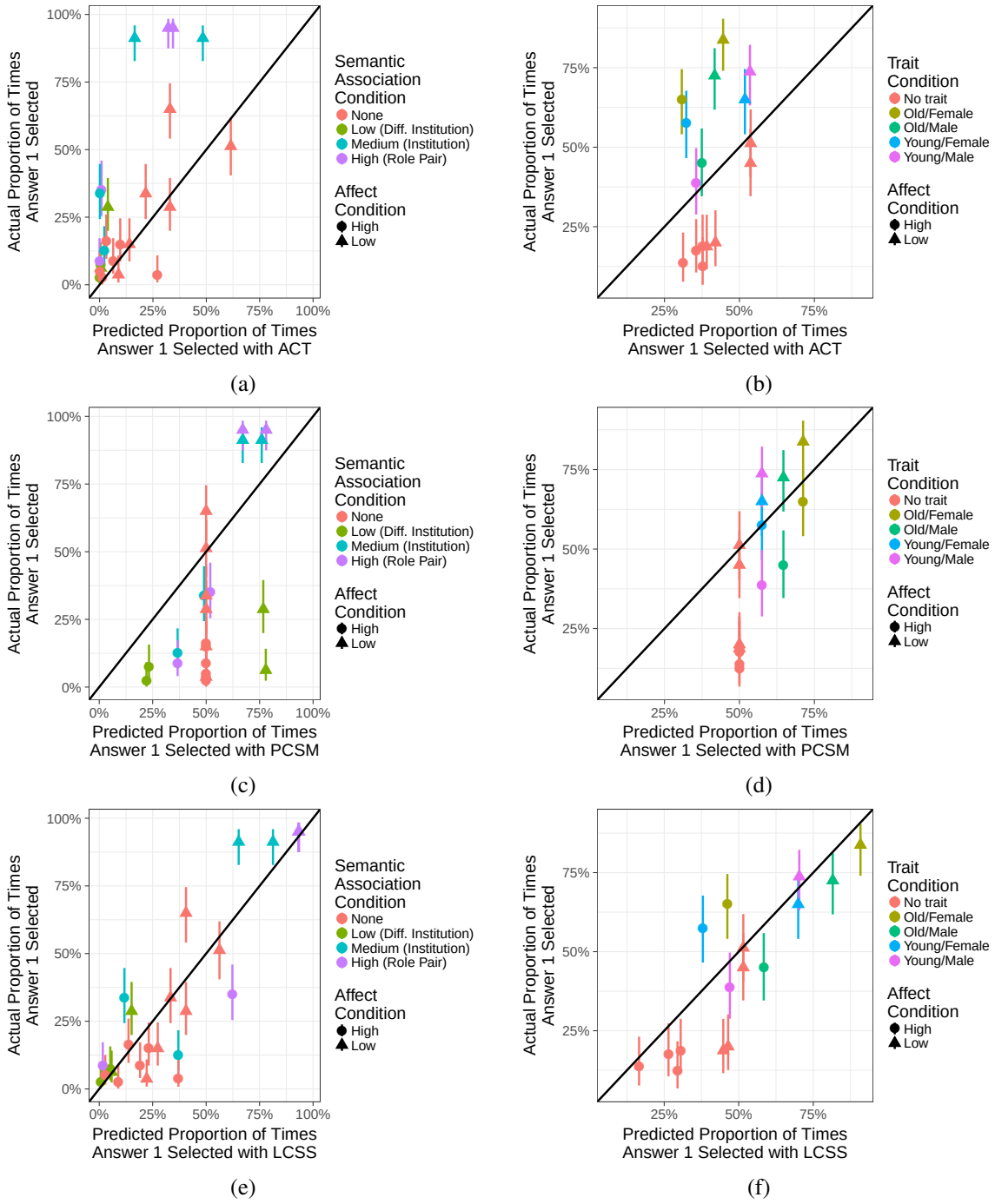

Figure 4: Comparisons of the predictions of the three models against empirical data.On each of the six subplots, the x-axis represents the predictions of a particular model for either the Affect vs. Association (left side) or the Affect vs. Trait (right) questions. The y-axis represents the value of $prop\_answer1$. One point on the plot is given for each survey question; these points also have Binomial confidence intervals (Agresti and Coull, 1998) along the y-axis to account for uncertainty in empirical estimates. Colors of the points give the cognitive conditions, which differ between the two types of questions. The shape of the points differentiate the high versus low affect condition. A diagonal line with a slope of one is plotted on each subplot; where the confidence interval of a question overlaps with this line, predictions made by a model are not significantly different than the empirical estimate of the "true" value of $prop\_answer1$. The top two subfigures ((a) and (b)) show results for ACT, the middle two ((c) and (d)) for PCSM, and the bottom two for LCSS ((e) and (f)).

out-predicts ACT by almost 60%, and PCSM by over 100%.

7 Discussion

The results of the previous section are somewhat unsurprising; considering both affective and cognitive information should intuitively be better than considering one or the other. Further, the somewhat involved mathematics of LCSS should not escape the very simple principles behind it. LCSS is grounded in the fact that ACT and PCS models provide important foundations for the identity labeling problem; it simply tries to combine their complementary advantages. What is surprising, then, is that until now, no existing work has combined these two strands of literature and model types that tangentially (in the case of ACT) or directly (in the case of PCS models) answer the identity labeling problem. And what is important is that in doing so, we address limitations of each existing model.

With respect to (Bayes)ACT and, in particular Heise and Mackinnon's (2010) extension of it, LCSS varies in one important way. Whereas Heise and MacKinnon view institutional constraints as being set a priori to limit the potential scope of identities available to individuals in a situation, LCSS assumes that identities and behaviors lie in a latent *semantic space* that, *in combination* with affective meanings, produces both how we label others and determines how we behave in social situations. With respect to identity labeling, the empirical data we present here suggests clearly that semantic information is not always used first, with affective information following behind. Instead, in line with modern understandings of cognition, "cognitive" and "affective" processes seem to interweave in producing the labeling decisions we see from survey respondents. Whether or not this is also the case for behavior remains to be seen, although our in-progress work suggests this to be the case as well (Morgan, 2018).

With respect to PCS models, LCSS addresses the fact that affective meanings of social behaviors are not easily encapsulated within a model of cognitive semantic networks. Instead of attempting to formalize how this might work in a network sense, we instead choose to adapt the notion of cognitive connections in networks into a latent space model that approximates similar phenomena. Practically, in turn, using a latent space representation of PCS models, as LCSS does, we are able to estimate a PCS-like model with significantly fewer parameters.[4]

However, it remains to be see whether or not this latent space, and the decision function LCSS proposes, can be estimated. Essentially, we have argued here that there is an unknown, potentially large number of dimensions of traits and semantic association dimensions that are important for identity labeling, and each adds a significant number of parameters to the decision model. Estimating these parameters will likely require novel, computational methods and large scale data. To this end, it is worth noting that while recent pushes in natural language processing have led to interesting new ways to characterize words and phrases (of which identities are of course a subset) into

[4]Whereas in a network, we would have to estimate the likelihood of a link between all pairs of concepts, here we can simply estimate positions in a finite-dimensional latent space.

latent dimensions (Mikolov et al., 2013), these dimensions are not individually interpretable, can be convoluted in their interpetability (Mimno and Thompson, 2017) and can given poor approximations when applied to questions of social behavior (Joseph and Carley, 2016). Developing similar methods that blend theoretically appealing notions of dimensions of meaning with computationally advanced algorithms is thus a potentially interesting avenue for solving the estimation issues in LCSS.

8 Conclusion

The present work makes three contributions to push us further in our understanding of the process by which we label ourselves and others with identities. First we propose and formalize the identity labeling problem, perhaps the most important subproblem in the study of the definition of situation. Second, we unite two strands of previously disparate existing literature on the identity labeling problem, one sociological and one psychological. We then show that existing predictive models from these two strands of literature are incapable of making correct predictions on a simplistic empirical dataset related to of the identity labeling problem. Finally, we show that by adopting the complementary strengths of these paradigms, we can develop a new model, which we title Latent Cognitive Social Spaces, that is better able to make such predictions.

Of course, the present work also harbors its fair share of limitations, both conceptually and methodologically. Methodologically, the way in which we test the proposed models does not at all reflect the identity labeling decisions that people face in their real life. It thus remains to be seen how various predictive models function, and must be amended for, "real world" settings. Additionally, the model we propose certainly less representative of cognition that a good deal of the work we draw upon. From a sociological perspective, however, such a trade off may be acceptable in return for theoretical consistency, generalizability and the relative parity that LCSS provides.

Conceptually, the present work assumes that an identity already holds a particular consensus meaning; we thus skirt any questions of how identities are constructed (Lamont, 2014). Similarly, we make simplistic assumptions about how identities are assigned, namely that we decide on one particular identity for an individual. This prevents us from delving deeper into multi-identity questions surrounding intersectionality and other related identity processes. It should be noted, however, that these conceptual issues are largely due to our desire to focus on a concrete predictive problem and to develop predictive models that address it. We believe that the theoretical concepts introduced by LCSS can in principle be extended to address questions of intersectionality and identity construction.

The present work also espouses a significant number of further questions. Most pressing in our opinion - with respect to the general model we have posed incorporating a consensus meaning space M and a decision function ϕ, what might we be able to say about a universal representation of identity meaning in M and how we use this to make a variety of decisions using different functions? That is, can we show that there exists a single common meaning

structure in which we interpret all identities and the context provided to them, and that it is only how we use this interpretation that varies? Is this space finite, and if so, how many dimensions are there (i.e. do we perceive a degree of "selfishness"?) Such conclusions would be hugely informative to discussions about how we construct and discuss stereotypes, both cognitive and affective, and how they impact social behavior.

References

Agresti, A. and Coull, B. A. (1998). Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician*, 52(2):119–126. ArticleType: research-article / Full publication date: May, 1998 / Copyright © 1998 American Statistical Association.

Ball, D. W. (1972). 'The Definition of Situation': Some Theoretical and Methodological Consequences of Taking WI Thomas Seriously. *Journal for the Theory of Social Behaviour*, 2(1):61–82.

Burke, P. J. (1980). The self: Measurement requirements from an interactionist perspective. *Social psychology quarterly*, pages 18–29.

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

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

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

Ehret, P. J., Monroe, B. M., and Read, S. J. (2014). Modeling the Dynamics of Evaluation A Multilevel Neural Network Implementation of the Iterative Reprocessing Model. *Personality and Social Psychology Review*, page 1088868314544221.

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.

Freeman, J. B. and Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological review*, 118(2):247.

Gaddis, S. M. (2017). How Black Are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies. *Sociological Science*, 4:469–489.

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

Heise, D. R. (2007). *Expressive Order*. Springer.

Heise, D. R. and MacKinnon, N. J. (2010). *Self, identity, and social institutions*. Palgrave Macmillan.

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

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

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.

Joseph, K. and Carley, K. M. (2016). Relating semantic similarity and semantic association to how humans label other people. *NLP+ CSS 2016*, page 1.

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

Lamont, M. (2014). Reflections inspired by Ethnic Boundary Making: Institutions, Power, Networks by Andreas Wimmer. *Ethnic and Racial Studies*, 37(5):814–819.

McFadden, D. (1980). Econometric models for probabilistic choice among products. *Journal of Business*, pages 13–29.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mimno, D. and Thompson, L. (2017). The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2873–2878.

Morgan, J. H. (2018). *The Duality of Identities and Groups: The Effects of Status Homophily on Social Interactions and Relations*. PhD thesis, Duke University.

Morgan, J. H., Rogers, K. B., and Hu, M. (2015). Distinguishing Normative Processes from Noise: A Comparison of Four Approaches to Modeling Impressions of Social Events. In Submission.

Penner, A. M. and Saperstein, A. (2013). Engendering Racial Perceptions An Intersectional Analysis of How Social Status Shapes Race. *Gender & Society*, 27(3):319–344.

Robinson, D. T. (2007). Control theories in sociology. *Annual Review of Sociology*, 33(1):157.

Rogers, K. B., SchrÃ¶der, T., and Scholl, W. (2013). The Affective Structure of Stereotype Content Behavior and Emotion in Intergroup Context. *Social Psychology Quarterly*, 76(2):125–150.

Schröder, T., Hoey, J., and Rogers, K. B. (2017). Modeling dynamic identities and uncertainty in social interactions: Bayesian affect control theory. *Am. Soc. Rev*.

Schröder, T., Rogers, K. B., Ike, S., Mell, J. N., and Scholl, W. (2013). Affective meanings of stereotyped social groups in cross-cultural comparison. *Group processes & intergroup relations*, page 1368430213491788.

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.

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.

Thomas, W. I. and Thomas, D. S. (1928). *The Child in America*. Knopf.

A Linking to BayesACT

The BayesACT model assumes that the deflection equation in Equation 13 is the logarithm of a probabilistic potential function (Equation 7 in Hoey et al., 2013)):

$$q(f', \tau') \propto \exp\left(-(f' - \tau')^T \Sigma^{-1} (f' - \tau')\right) \tag{13}$$

In this model, the fundamental is defined as f' and the transient is defined as τ'. This is done in order to emphasize that in BayesACT, unlike in ACT, the fundamental and the transient are (time varying) probability distributions. Because it assumes a probabilistic form, the BayesACT model thus can thus directly give a probability that a given identity is the "correct" identity for a given situation. Relevant to the present work is that the form of Equation 13 is clearly similar to the form of Equation 2; indeed, Equation 2 is simply a normalized version of the model posed by Hoey et al. (2013) where we also integrate over f' and τ':

$$p(i_y|M, r) = p(i_y|f', \tau') \propto \frac{\int_{f',\tau'} q(f'_{i_y}, \tau'_{i_y})}{\sum_{k \in I} \int_{f',\tau'} q(f'_{i_y}, \tau'_{i_y})} \tag{14}$$

This is a discrete choice model in the form of Equation 2 where $\phi(i_a, M, r) = \int_{f',\tau'} \log(q(f'_{i_y}, \tau'_{i_y}))$. In Appendix B of Hoey et al. (2013), the authors demonstrate how BayesACT reduces to the same predictions as the original ACT

model under certain conditions. Because computing the integration over f' and τ' by either derivational or numerical approaches would complicate the simplicity of the mathematical argument posed in the present work, we will therefore assume that it is under these conditions that the predictions are made here. In other words, we use predictions from the "original" version of ACT described above, leaving a comparison with predictions made by BayesACT under different parameter settings to future work.

One final point worth noting on BayesACT is that the theory, unlike ACT, requires no assumption that identities are known for social interaction to take place. Instead, BayesACT only requires a *distribution* over identities is known for the actor and object. The present work, instead, focuses on data and methods that imply people often choose a label for actors and object. It should be noted, however, that the more general definition of the identity labeling problem proposed in Section 5 requires only this same probability distribution considered by BayesACT; the final selection is necessary to compare with our survey results but not in principle. In this way, the formalizations of the identity labeling problem are exactly the same as those embraced by BayesACT.