A New Approach for Discovering Social Beliefs about Ethnic Structure *

Thomas B. Pepinsky
Department of Government
Cornell University
pepinsky@cornell.edu

First draft: June 4, 2017
This draft: April 11, 2018

Abstract

This paper introduces a new approach to discovering and exploring society-wide social beliefs about ethnic structure. Rooted in computational text analysis, it combines the strengths of both qualitative and survey-based approaches to the study of ethnicity. I use a structural topic modeling approach (Roberts, Stewart, and Airoldi 2016) to uncover general patterns in found in open-ended questions about ethnic groups. I then predict which patterns of responses tend to be associated with which ethnic groups, and use information about survey respondents to explore heterogeneity in social beliefs. To illustrate the method at work, I use original survey data from Malaysia to understand social beliefs about Malay, Chinese, Indian, and Arab ethnic groups in Malaysia. I also pair data on Malays in Malaysia with original survey data on Malays in Sumatra, Indonesia, to explore cross-national differences in social beliefs about Malays.

*Thanks to Kevin Foley, David Laitin, Rich Nielsen, and participants at the Elections and Participation Conference at National Chengchi University for helpful comments on a preliminary draft. Thanks also to Dodi Ambardi, Azlyna Halit, Hendro Prasetyo, Ben Suffian, and Tan Seng Keat for discussions and comments.

A New Approach for Discovering Social Beliefs about Ethnic Structure

1 Introduction

Contemporary research on ethnic politics largely agrees that ethnic categories are socially constructed. However, although social beliefs about ethnicity and ethnic identity are not based on objective facts about groups or individuals, they remain socially meaningful and politically consequential. It remains difficult to characterize these society-wide social beliefs about differences across ethnic groups. One general approach to characterizing social beliefs about ethnic structure is to focus carefully on text, speech, the media, or the researcher's own personal experience. This risks selection bias by focusing on nonrepresentative subgroups from the population of interest. Another approach, asking survey respondents to respond to pre-specified questions about ethnicity, risks confirmation bias by supplying the beliefs to the respondents *ex ante* rather than allowing respondents' own beliefs to emerge.

This paper introduces a new approach to discovering and exploring society-wide social beliefs about ethnic structure that combines the strengths of both qualitative and survey-based approaches to the study of ethnicity. The essence of this new approach is simple: in a representative sample drawn from the population of interest, ask respondents what are the words that come to mind when they think about particular ethnic groups. Treating these open-ended survey responses as data, then estimate the general patterns in topical content across these responses, and predict which patterns of responses tend to be associated with which ethnic groups. Information about survey respondents can be used to uncover heterogeneity in social beliefs within the population under study. To illustrate the method at work, I use original survey data from Malaysia, collected in 2017, to understand social beliefs about Malay, Chinese, Indian, and Arab

ethnic groups in Malaysia.

Relative to existing approaches to studying social beliefs about ethnic structure, the approach introduced here has five attractive features. First, it is *general*, allowing researchers to characterize society-wide beliefs about ethnic structure without relying on personal experiences that may not be representative of society at large. Second, it is *probabilistic*, allowing for uncertainty, disagreement, and contestation of social beliefs about ethnic structure. Third, it is *structural*, designed to explore not just differences across ethnic groups but also explanations for why different groups within society characterize ethnic groups differently, based on a microfounded theory of survey response. Four, it is *non-exclusive*, meaning that society-wide social beliefs may be common to multiple ethnic groups in ways that existing approaches find difficult to capture (see Pepinsky, Liddle, and Mujani 2018). Finally, it is easy to implement in a standard survey design.

Characterizing social beliefs about ethnic structure contributes more broadly to the study of identity, ethnic conflict, public goods, and other areas of research in political science, economics, sociology, and related disciplines in which differences among groups may affect individual behavior or social outcomes. The conceptual framework outlined below allows for the comparative study of ethnic structure across groups and societies by explicitly modeling social beliefs as individual beliefs which may be uncovered using survey-based methods. Moreover, many models of ethnic conflict allow for the potential for conflict to vary as a function of "group differences" (see Esteban, Mayoral, and Ray 2012) but group differences are poorly measured in applied empirical work. The most common way to capture these differences is to examine linguistic distances across groups (Fearon 2003), which reduces group differences to a single dimension that is unlikely to capture socially meaningful distinctions across populations. The approach here may be used to develop a research program that characterizes the magnitude, dimensionality, and meaning of group differences using categories that are socially relevant to the groups under study. Finally, most quantitative studies of ethnic conflict ignore the content of group identities, or invoke natural experiments in which cultural differences are held constant (Posner 2004), and instead focus

on the size and/or distribution of groups within a society in order to understand the effects of ethnic politics. Studying social beliefs about ethnic structure extends this powerful line of inquiry to ask not just how many groups there are in society and how are they distributed, but what do their members believe about one another.

2 Conceptual Framework

A society is comprised of individuals who may be divided into social groups. These social groups serve as a reference category for individuals' own individual identities, but they also provide meaning and structure to social interactions. The conceptual framework outlined here focuses on ethnic groups with the goal of understanding society-wide ethnic structure, but through analogous reasoning may be applied to religion, class, race, or any other socially meaningful identity category.

Beliefs refer to individuals' understandings of the nature of the world around them. These include ontologies, truth claims, normative values, possibilities for action, and others (see Goldstein and Keohane 1993, pp. 7-8). An example of a belief about an ethnic group is "members of ethnic group A are lazy," which combines an ontological claim about the existence of a group ("there exists an ethnic group A") and a normatively valued truth claim associated with that group ("laziness is a property of the members of A"). *Social beliefs* refer to understandings about the social world that are common across groups of individuals. Social beliefs about ethnic groups are consequential because they represent general patterns of beliefs across multiple individuals, and as such, they shape individual and group behavior. If the statements "members of ethnic group A are lazy" and "members of ethnic group B are industrious" are both social beliefs, then it will follow, for example, that an individual faced with a hiring decision and seeking to maximize employee effort may prefer to hire from group B over group A. Likewise, a social planner seeking to eradicate intergroup wealth disparities may encourage effort from group A and/or discourage

effort from group B.

The relationship between individual and social beliefs is both causal and constitutive. Beliefs *are* social beliefs when they are widely shared across individuals; this is a constitutive relationship in which social beliefs may be exhaustively disaggregated into the beliefs of individuals. Individuals may also come to hold their beliefs *because* they are widely shared among other individuals; this is a causal relationship in which socialization determines individual beliefs. In either case, however, it is possible to identify social beliefs at any one point in time by identifying patterns in individual beliefs that are shared across groups. I use the term *ethnic structure* to denote the collection of social beliefs about ethnic groups in a given society. It denotes both the relevant groups in society and the social beliefs about each of them.[1]

Crucially, social beliefs about ethnicity are beliefs, not facts. This grounds the study of ethnic structure in a social constructivist approach to ethnic identity (Chandra 2012). Also crucially, social beliefs are common, not unanimous. A challenge for constructivist approaches to ethnic identity how to reconcile widely-held beliefs with difference, contestation, and opposition to these beliefs. By conceptualizing social beliefs as nothing more than the aggregation of commonly held individual beliefs, it is straightforward to see how individuals may hold different beliefs than those around them—social beliefs still exist even if individuals vary. Likewise, groups within society may collectively hold different views about themselves than the rest of society. The belief "members of ethnic group A are lazy" may be a social belief for group B but not group A itself.

There are three common ways that social scientists seek to understand ethnic structure. The first approach is to ignore social beliefs about ethnic structure entirely, and to focus merely on the number, size, and/or distribution of ethnic groups. This is the case in common indices of ethnic heterogeneity such as the ethnolinguistic fractionalization index (Taylor and Hudson 1972). This approach is appropriate if social beliefs follow entirely from ethnic demography, or if social

[1]Below, I will focus on discovering social beliefs rather than the identity of the relevant groups in society. Future work may extend the approach outlined here with the goal of discovering what groups in society are believed to be relevant.

beliefs about ethnic structure are irrelevant. However, if the research objective is to understand social beliefs, or to investigate their causal effects, or if non-demographic factors such as historical legacies, public policy, or lived experience shape these social beliefs, then an alternative is required.

One such alternative approach is ethnographic and contextual: using historical texts, personal experience, and/or ethnographic data to understand the social meanings behind identity categories (Brubaker et al. 2006; Nagata 1974). This approach is particularly useful for allowing researchers to uncover social meanings, nuance, and particularities of ethnic identities and how they are understood. However, such approaches is subject to sampling bias, and it is difficult to generalize beyond local experiences or particular texts and contexts to the broader target population.

A second alternative approach leverages public opinion data to characterize more generally the pattern of responses to questions about ethnic identity across the population of interest. Such approaches may ask respondents how strongly they agree or disagree with a set of statements about ethnic groups, or to rate ethnic groups on a set of attributes. For example, the General Social Survey in the United States frequently asks respondents their beliefs about American racial groups, such as their "commitment to strong families" or whether they are "Hard working" or "lazy" (Smith et al. 2016). Although patterns of responses are more easily generalizable, they require the relevant beliefs about ethnic structure to be known *ex ante* by the researcher, and it is difficult to weight various responses according to how meaningful they are. Respondents may provide answers to questions that are irrelevant, and not be given the opportunity to answer questions that are relevant. Even worse, relevance itself may vary across meaningful subgroups within the population.

The innovation I propose here preserves the representativeness of survey-based approaches without imposing researcher-derived categories or priorities on ethnic structure. To conceptualize how to characterize social beliefs from survey data, it is helpful to represent individual survey

responses as reflecting both social and individual beliefs about ethnic groups. An individual i's beliefs B about ethnic group k depends on society-wide social beliefs about group k, beliefs about group k that are common across demographic subgroups j of which i is a member, and i's idiosyncratic individual-level beliefs about ethnic group k:

$$B_{ijk} = f(EthGroup_k, Demo_j, Individual_i) \tag{1}$$

Responses from demographic group j differ across ethnic groups k, and that difference can be decomposed into the typical response that demographic group j provides and those that it specifically gives to ethnic group k: $\delta_j + \beta EthGroup_{jk}$. Social beliefs about ethnic groups k can be likewise decomposed into typical beliefs for shared about all ethnic groups and those specific to each ethnic groups $\alpha + \gamma EthGroup_k$. Combining these expressions under a linear specification relating societal, group, and individual-level responses results in the following generic model of social beliefs about K ethnic groups from J demographic groups.

$$B_{ijk} = \alpha_{ijk} + \sum_{j=1}^{J} \delta_j Demo_j + \sum_{k=1}^{K} \gamma_k EthGroup_k + \sum_{j=1}^{J} \sum_{k=1}^{K} \beta_{jk}(EthGroup_k \cdot Demo_j) + e_{ijk} \tag{2}$$

Expression 2 represents social beliefs as a function of the ethnic groups, respondents' demographic characteristics, the pairwise interactions between ethnic groups and respondents' demographic characteristics, and idiosyncratic respondent-level factors captured by the term e_{ijk}.

What remains is the task of inferring beliefs B_{ijk} from individual survey responses. The approach I follow here draws on recent advances in computational text analysis to estimate respondent beliefs from open-ended survey responses. Rather than provide respondents with a series of structured questions about ethnic groups and their views and opinions about them, I instead invite respondents to respond freely to an open-ended question to describe what comes to mind when they think about different ethnic groups. The logic underlying this approach is simple: that

respondents will name only those beliefs that are salient to them, and general patterns will emerge from the collection of responses in aggregate. This survey-based approach is easy to implement, and the context can generate a natural response without heightening respondents' awareness of researchers' concerns or priorities, which may heighten the risk of social desirability bias or related concerns.

Given a collection of survey responses about ethnic groups, the next problem is uncover respondent beliefs from them. Topic models (Blei 2012) are well-suited to this task. Topic models uncover patterns in textual data, called *topics* that in this case correspond to beliefs, across *documents*, which in this case correspond to survey responses. However, research goal here is to identify not just the general beliefs that responents have about ethnic groups, but how these beliefs differ accross ethnic groups and respondents. Conceptually, this is a model of predicting topics: when a respondent of type j describes ethnic group k, what topics are more or less likely? Structural topic models (Roberts, Stewart, and Airoldi 2016) that jointly model topics and document-level covariates provide a natural framework for accomplishing this task. Roberts et al. 2014 have shown how structural topic models can be used to analyze open-ended survey responses, and I adapt their framework to uncover social beliefs across survey responses.

3 Data and Methods

To demonstrate how my approach can discover social beliefs about ethnic structure from survey data, I focus on Malaysia as a case study. Malaysia is a middle-income country with a population of approximately 30 million people located in Southeast Asia, with territory divided between the Malay peninsula and the northern coast of the island of Borneo (shared with Indonesia and Brunei Darussalam). Malaysia is also a plural society: the majority of the country's population identifies as Malay, with the remainder of the population identifying as ethnic Chinese, ethnic Indian, or as a member of an additional set of non-Malay indigenous groups, most of whom live in

Malaysian Borneo. Together, Malays and non-Malay indigenous groups are termed *bumiputera*, or "sons of the soil." Official statistics from 2016 place the *bumiputera* population share at 68.6%, with the remainder divided among Chinese (23.5%), Indian (7.0%), and other (1.0%) (Department of Statistics Malaysia 2016). Overlaying this official breakdown of the Malaysian population are other migrant identities such as Arab, Javanese, Portuguese, and others; generations of intermarriage results in a society where many people who identify as Malay in one situation may identify as Arab or Javanese in another (Nagata 1974).

As a plural society, ethnic identity in Malaysia has long been associated with particular social beliefs. Malays have widely been associated with laziness or unwillingness to work, a belief that dates to colonial times (Alatas 1977) but which remains current in Malaysian politics, even among Malay politicians (see Mohamad 2003, p. 155). Malayness is also a political project: the Malaysian constitution defines "Malay" with reference to language, custom, and Islam ("Federal Constitution of Malaysia", Article 160). Chinese Malaysians are commonly viewed as being industrious, or less charitably, greedy. Indians are commonly associated with stereotypes of drunkenness and crime. Ethnic chauvinism is thus rampant in Malaysia, but the building of a trans- or post-ethnic Malaysian identity remains a high priority for those who oppose Malaysia's current political system, in which most large parties are defined by ethnicity or religion rather than by programmatic platform. Moreover, social beliefs surely differ across demographic groups. As such, it is not clear whether the set of social beliefs listed above reflect actual social beliefs in Malaysia or, perhaps, one researcher's own biases or non-representative experiences. For these reasons—identity is politically and socially salient but also contested across multiple dimensions—Malaysia is an idea laboratory in which to explore social beliefs about ethnic structure.

As part of a larger effort to understand ethnic politics and identity in Malaysia, 1200 Malaysians were surveyed in March 2017 (details about the survey itself are available in Appendix S1). In addition to a series of demographic questions, respondents were asked to respond to the following prompt.

> Now I am going to ask you about several different ethnic groups. I would like to you to tell me two (2) things that come to mind when you think about these groups. There is no right or wrong answer; you may think of particular words or phrases, or perhaps nothing at all.
>
> 1. Malays
> 2. Chinese
> 3. Arabs
> 4. Indians

In addition to the trichotomous Malay-Chinese-Indian distinction, the survey also included Arabs in order to capture an important identity category in Malaysia that deliberately does not fit into the official narrative of Malaysian identity, and which can reveal subtleties in the ways in which partially overlapping identity categories are understood. Responses were recorded both in Malay (*Bahasa Melayu*, the national language of Malaysia) and English. In what follows, data are analyzed from the Malay language responses; for illustrative purposes, however, English-language translations are provided.

To prepare the data for processing, each pair of responses was merged by ethnic group, then all responses were stacked across ethnic groups by respondent. Next, because the Malay language features a complex system of morphological affixes, I use the stemming algorithm of Nazief and Adriani 1996 as implemented by Setiabudi 2017 for the Indonesian language to replace each affixed word with its semantic root. Standard Indonesian and standard Malay are two standardized versions of the same Malay-based trade language, and share an identical set of morphological features (Collins 1998),[2] so stemming methods developed for Indonesian work equally well for Malay. For example, a word such as *berpendidikan* [= educated] is stemmed as *ber- pen-* ***didik*** *-an*, with *didik* [= learn] as the semantic root.[3] Stemming increases comparability across survey

[2]The exceptions are only found in informal registers of spoken Indonesian and Malay.
[3]See Adriani et al. 2007 for a general review of stemming in Indonesian.

Table 1: Data for Analysis

R No.	Description	Group (EGOI)	Age Cat.	Ethnicity (EGOR)
1	our nation malay language	Malay	5	Malay
1	greedy suppressing	Chinese	5	Malay
1	...	...	5	Malay
2	islam	Malay	7	Malay
2	immigrant power over malays	Chinese	7	Malay
2	...	...	7	Malay

Table 2: Demographic Covariates

Variable	Description
EGOR	Categorical variable recording respondent's ethnic group
Gender	1 = Respondent is female, 0 otherwise
Age	Ordinal variable recording respondent's age, 9 values
Peninsula	1 = Respondent lives in Peninsular Malaysia, 0 otherwise

responses by allowing *berpendidikan* to be analyzed as equivalent to other, related words used by respondents such as *didikan* and *pendidikan* [= education].[4]

The result is illustrated in Table 1. The variable *Description* contains each respondent's responses for each group. To distinguish between the ethnic identity of interest and the ethnic identity of the survey respondent, which may also be of theoretical interest, the variable *EGOI* corresponds to the Ethnic Group of Interest being described in *Description*, whereas the variable *EGOR* records the Ethnic Group of the Respondent (recorded as *Ethnicity* in Table **??**).

4 Results

Using these data, I fit a structural topic model in which the prevalence prior is set by four socially meaningful demographic covariates, each interacted with the categorical *EGOI* to allow for ethnic group-specific differences in priors across these covariates. A description of these four demographic covariates appears in Table 2. Structural topic models, like all topic models, require

[4]The stemmer is "aggressive" in that it returns the semantic root of every word it encounters. There is some risk of overstemming in the case of figurative words such as *pengangguran* [= unemployment, fig. "wine-ness"], stemmed to *anggur* [= grape/wine], but such cases are rare.

the analyst to specify the number of topics that exist in the data. The "correct" number of topics depends on the research question at hand: using fewer topics highlights coarse differences at the expense of nuance, whereas using more topics may reduce the semantic coherence of any one topic. In preliminary explorations I found that a structural topic model with twenty topics provides an acceptable balance of nuance and coherence. For purposes of comparison, results from topic models with ten and and thirty topics are presented in Appendix S2.

Table 3 displays each of the twenty topics obtained from the structural topic model, alongside the words identified as having the highest probability of appearing in each topic. As in any

Table 3: Topics

Topic	Highest Probability Words (Malay)	Highest Probability Words (English)
1	orang bahasa arab budi tutur besar asing	person language arab reason express large foreign
2	agama mereka raja kepada pegang jaga gantung	religion them king to hold protect depend
3	banyak ekonomi mata putih jahat sepet sangat	many economy eye white evil slanted very
4	lebih sama nasi dari buka roti hadap	more same rice than open bread before
5	satu hidup maju tolong tiada tipu amah	one life develop help is none trick servant
6	diri pandai sendiri penting sombong tetap kawan	self smart alone important arrogant remain friend
7	dengan lain gaul berbeza didik sosial masalah	with other mix different learn social problem
8	suka susah daripada senang gaduh masa ambil	like difficult than like noisy time take
9	bangsa kuasa cara minoriti minum pelbagai arak	nation power way minority drink very liquor
10	baik tidak hati boleh percaya asa putus	good no heart may believe hope decide
11	kurang makan tinggi dengki tudung cepat gangster	less eat high jealous headscarf fast gangster
12	malas sopan santun muslim pakai dapat cantik	lazy proper mannered muslim wear receive pretty
13	kaum kulit cakap gelap india hitam hindu	group skin speak dark india black hindu
14	layu malaysian kasar tak sahaja halal sifat	withered malaysian coarse not just halal characteristic
15	rajin buat buddha berdikari hidung bagus comel	industrious do buddha self-standing nose good appealing
16	niaga kerja kaya duit usaha bijak cari	business work rich pay effort policy search
17	islam kuat mesra baju khinzir bukan jubah	islam strong friendly clothes blouse pig not robe
18	dan mudah budaya lupa sikap usaha bantu	and easy culture forget attitude work help
19	malaysia dalam cina untuk adat mahir majoriti	malaysia in china for culture master majority
20	yang ada negara pemikiran tahu baikada laku	that is state thought know even behave

application of text-based methods, the structural topical model classifies words into topics but does not supply them with semantic content, which must be inferred by the analyst. Inspecting the elements of Table 3, however, topical content is readily apparent. Topic 1, for example—which includes the words "person language arab reason express large foreign"—evidently focuses on language foreignness. Topic 2—"religion them king to hold protect depend"—covers politics and religion. Topic 13—"group skin speak dark india black hindu"—addresses physical appearance.

Topic 16—"business work rich pay effort policy search"—covers economic activity. These topics, estimated from the aggregated survey responses of 1200 Malaysians describing four different ethnic groups, reveal coherent ways in which Malaysian think about differences across ethnic groups. In Appendix S3 I provide examples drawn from texts that represent each of these four topics.

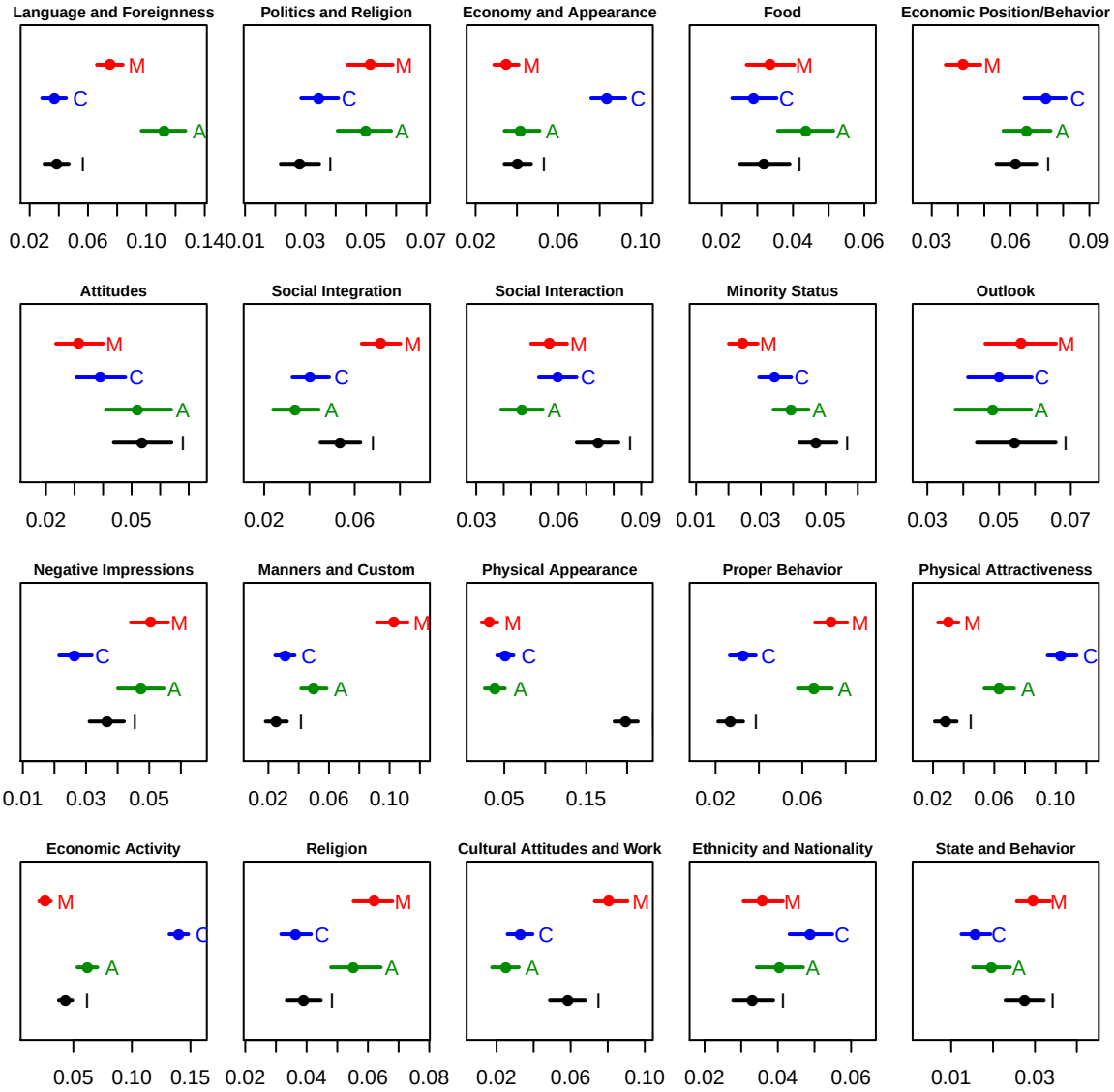
To structure these beliefs about ethnic differences further, I use the categories of *EGOI* in interaction with the demographic covariates in Table 2 to predict the prevalence of each topic in the survey responses. Writing $\mathbf{B}_t$ as a vector of $t = 1, 2, \ldots T$ social beliefs B_{ijk} (with subscripts i indexing individuals, j indexing demographic characteristics from Table 2, and k indexing *EGOI*), and stacking expression 2, we have

$$\mathbf{B}_t = \boldsymbol{\alpha}_t + \sum_{j=1}^{J} \boldsymbol{\delta}_{jt} Demo_j + \sum_{k=1}^{K} \boldsymbol{\gamma}_{kt} EGOI_k + \sum_{j=1}^{J} \sum_{k=1}^{K} \boldsymbol{\beta}_{jkt} (EGOI_k \cdot Demo_j) + \mathbf{e}_t \tag{3}$$

The parameters of interest in Expression 3 are $\boldsymbol{\gamma}_{kt}$, a $T \times K$ matrix of parameters which measure the strength of the association between ethnic group k and belief t, and $\boldsymbol{\beta}_{jkt}$, $T \times (K \times J)$ matrix of parameters that capture differences in social beliefs across demographic groups j. This procedure allows me to associate topics with ethnic groups; to answer the question of "what social beliefs are typically associated with which ethnic groups?"

Figure 1 visualizes these results by plotting the predicted prevalence of each of the twenty topics for each of the four ethnic groups (that is, the categories of *EGOI*). In these calculations, all other covariates are held at their sample medians. Topics 1-20 are labeled using intuitive summaries of the topical content as revealed in Table 3. Figure 1, for example, shows that the topic of Language and Foreignness (Topic 1) is more likely to be used in reference to Arabs—and to a lesser extent Malays—relative to Chinese and Indians. The topic Politics and Religion (Topic 2) is more likely to be associated with Malays and Arabs than Chinese and Indians. By contrast, Physical Appearance (Topic 13) is almost always associated with Indians. Economic Activity

Figure 1: Topic Prevalence



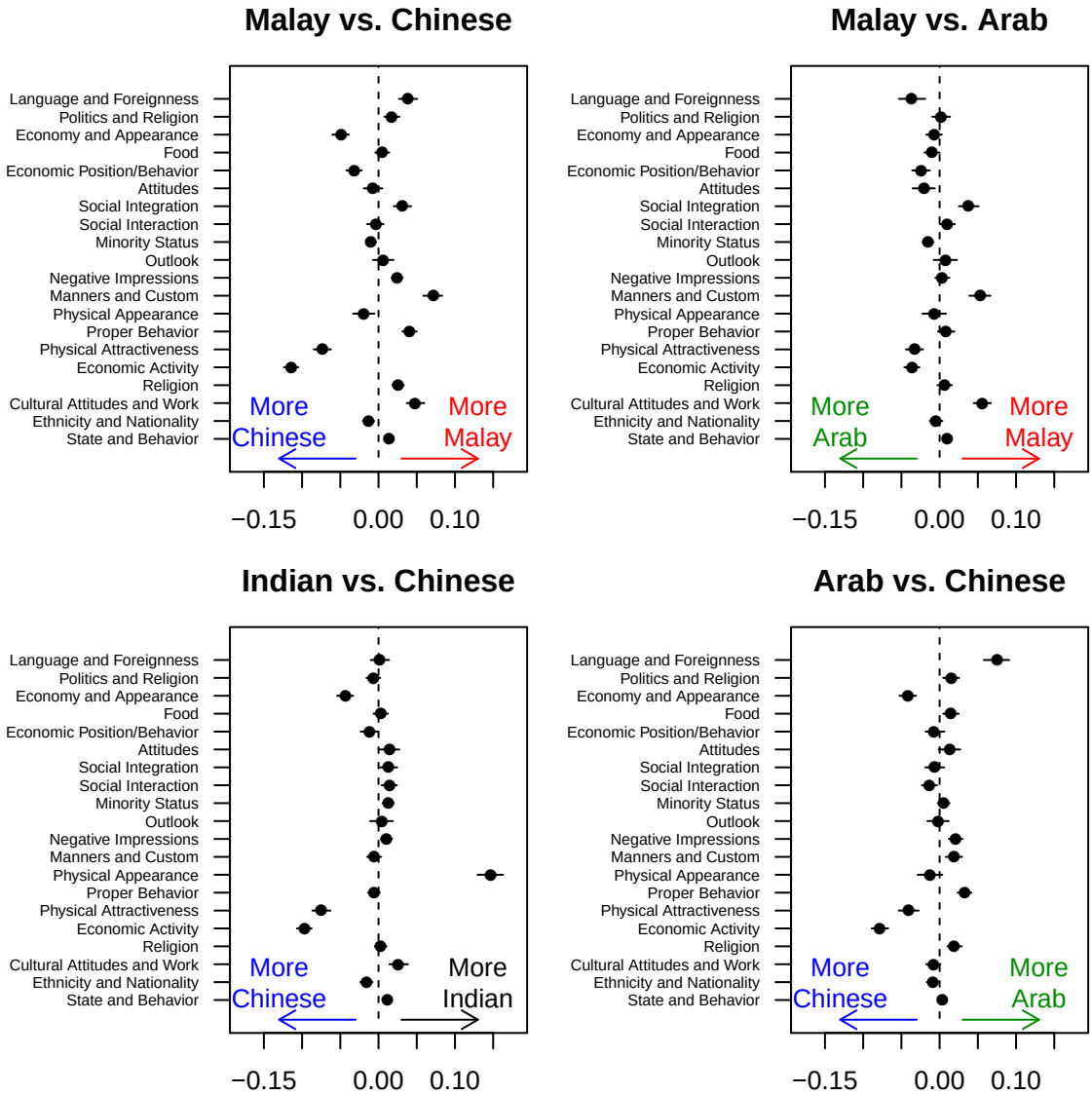
Note: Each figure plots the predicted prevalence (and its 95% confidence interval) of each of the twenty topics estimated from *Description* for each ethnic group. X-axis scales vary across topics. M=Malay, C=Chinese, A=Arab, I=Indian.

(Topic 16) is associated primarily with Chinese, and secondarily with Arabs.

An alternative way to visualize these same results is to compare the prevalence of each topic across pairs of ethnic groups. This more directly answers the question "what social beliefs are generally more associated with groups A versus B?" Figure 2 presents four pairs of ethnic groups:

Malays versus Chinese, Malays versus Arabs, Indians versus Chinese, and Arabs versus Chinese. Direct comparison highlights a number of meaningful differences in the social meanings associ-

Figure 2: Four Comparisons



Note: Each figure plots the difference in predicted prevalence (and its 95% confidence interval) of each of the twenty topics estimated from *Description* for a selected pair of ethnic groups.

ated with different Malaysian ethnic groups. As might be expected, Topics 15 and 16 covering economic activity and physical attractiveness are more strongly associated with Chinese than

Malays, whereas Topic 12—covering manners, custom, and Muslim identity—more with Malays. Topic 12 *also* distinguishes Malays from Arabs, along with Topics 7 (on social difference and mixing) and 18 (on cultural attitudes and work). Topics 15 and 16 covering economic activity and physical attractiveness distinguish Chinese from Indians, whereas skin color and physical appearance are more associated with Indians. Although Topic 16 is more commonly applied to Arabs versus Malays and Chinese versus Malays, it is also more commonly invoked for Chinese than Arabs.

We can identify those topics with the highest prevalence score by ethnic group, which will generally identify which social beliefs are most generally associated with each ethnic group. Table 4 lists the top three topics associated with each of the four groups described in the survey: These results illustrate the very different social beliefs that Malaysians have about different eth-

Table 4: Top Three Topics

Malays		**Chinese**	
Topic	Description	Topic	Description
12	Manners and Custom	16	Economic Activity
18	Cultural Attitudes and Work	15	Physical Attractiveness
1	Language and Foreignness	3	Economy and Appearance

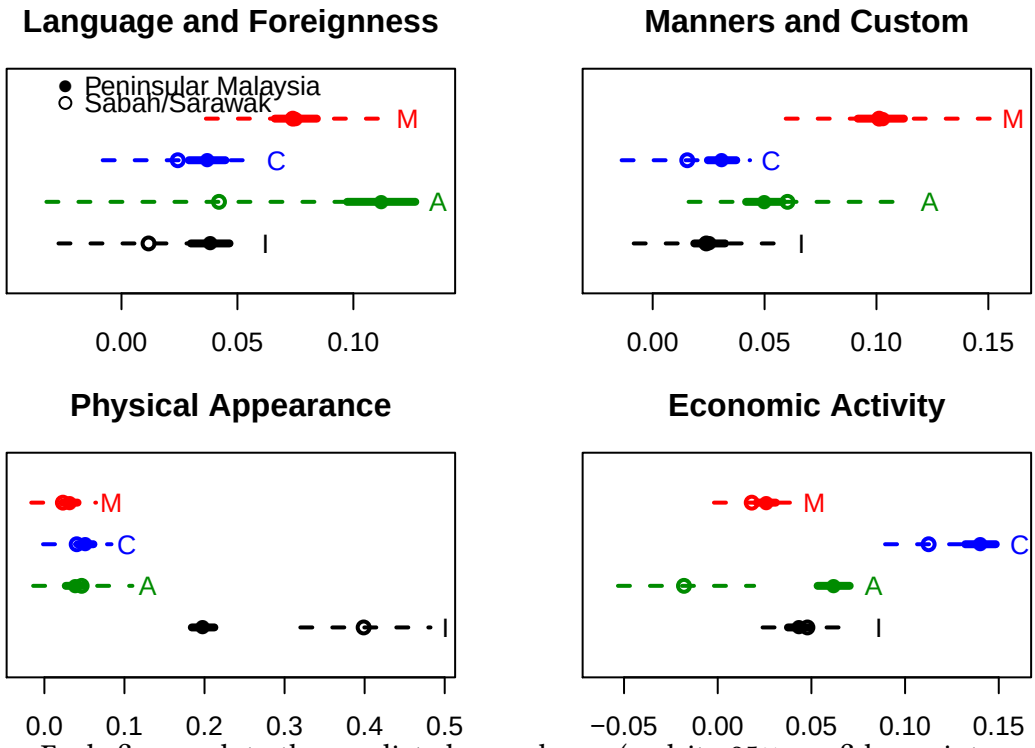
Arabs		**Indians**	
Topic	Description	Topic	Description
1	Language and Foreignness	13	Physical Appearance
14	Behavior and Islam	8	Social Behavior
16	Economic Activity	5	Economic Position/Behavior

nic groups. Malayness is associated with manners, custom, Islam, and language; Arabness is associated with language too but more strongly with Islam and economic activity. Beliefs about Chineseness focus on economic position and behavior as well as physical appearance—with specific mentions of skin color, eye shape, and nose. Beliefs about Indianness also focus on physical appearance—that is, skin color—but also common stereotypes about social behavior. Malaysia's ethnic structure thus differentiates groups according to religion, custom, and language versus

physical appearance, economic function, and social position.

We can also explore heterogeneity in social beliefs across Malaysians by investigating the interactions between respondents' demographic characteristics and topic prevalence across ethnic groups. One important social cleavage in Malaysian society is between peninsular Malaysia and the Malaysian states of Sabah and Sarawak located on the island of Borneo. With a different colonial history and unique social and ethnic structure, it is widely understood that Sabah and Sarawak together present an alternative ethnic structure than that found in peninsular Malaysia. Figure 3 investigates whether topic prevalence differs across respondents (of any ethnic group) who reside in peninsular Malaysia versus Sabah or Sarawak. The focus here is on the most prevalent topics for each ethnic group (see Table 4). Contrary to expectations, there is little

Figure 3: Four Comparisons

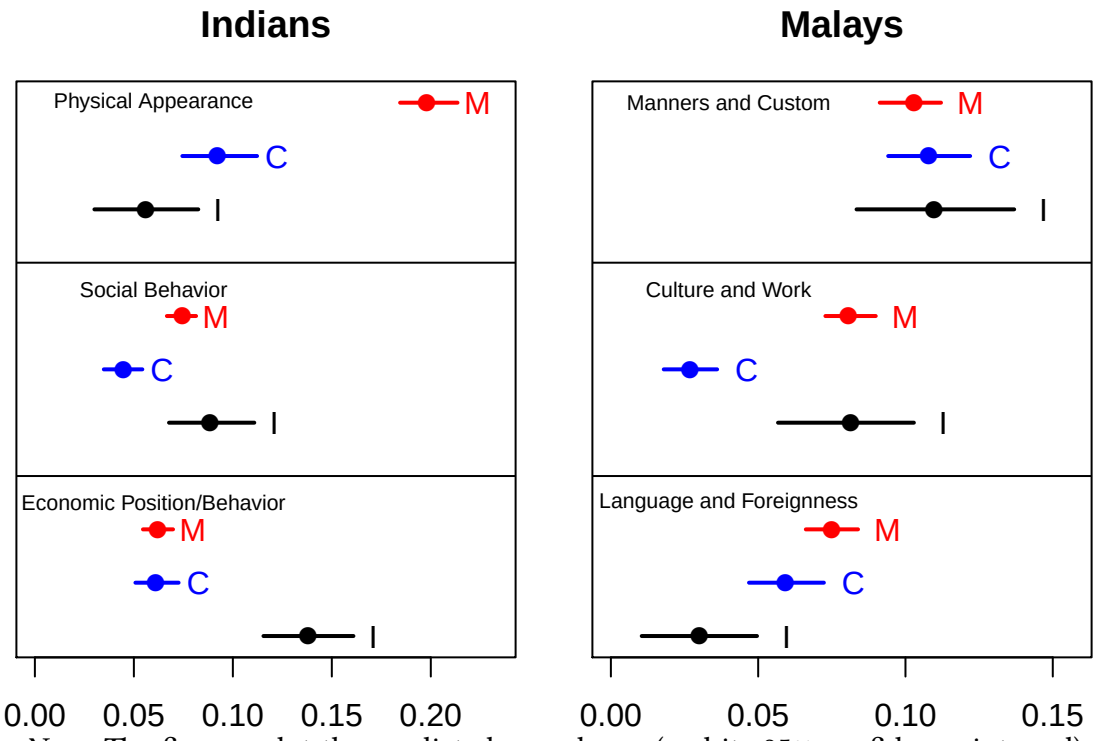


Note: Each figure plots the predicted prevalence (and its 95% confidence interval) of topics estimated from *Description*, comparing respondents from peninsular Malaysia versus Sabah and Sarawak.

evidence that social beliefs about ethnic groups differ between peninsular Malaysians and Sabahans/Sarawakians. In many instances the point estimates topic prevalence differs substantially, but the wide confidence intervals around the estimates for Sabah/Sarawak prevent us from concluding that these differences are statistically significant. The one clear exception is the association of Arabs with economic activity (Topic 16), which appears to be solely a product of peninsular Malaysian respondents.

It might also be the case that social beliefs about ethnic structure vary by ethnic group. Consider the striking results above in Figure 1 about Indians in Malaysia, strongly associated with physical appearance. It is useful to distinguish whether Indian Malaysians themselves associate Indianness with physical appearance, and if not, what other characteristic is dominant. To do this, Figure 4 looks at the three topics most strongly associated with Indians from Table 4, and for each, estimates plots prevalence among Malay, Chinese, and Indian respondents. The second panel in Figure 4 does the same for the three topics most strong associate with Malays. The left hand figure, which focuses on social beliefs about Indian Malaysians, reveals that the aggregate picture of Malaysian social beliefs associating Indians with a particular physical appearance is driven by Malay respondents, who are the plurality of respondents in the survey sample. Indians themselves are more likely to associate Indianness with social behaviors. The right hand figure, by contrast, shows that there is very little difference among Malays, Chinese, and Indians in association Malayness with manners, custom, and Muslim identity (Topic 12). There is more heterogeneity across Malays and other respondents in the other two topics. In this way, a text-as-data approach to studying ethnic identity can characterized precisely how beliefs about ethnic structure vary. Such tools can be used to better understand the subtleties in prejudice and bias that are not generally found across all members of society.

Figure 4: Differences by Ethnic Group



Note: The figures plot the predicted prevalence (and its 95% confidence interval) of the three topics most strongly associated with Indians and Malays, comparing across Malay, Chinese, and Indian respondents.

5 Cross-National Comparisons

The discussion so far has focused exclusively on discovering social beliefs about ethnic structure within one country. It is also possible, however, to use this same approach to uncover differences in social beliefs across countries. As a final exercise, I illustrate how this is possible using an additional round of survey data collected in three provinces and one city in Sumatra. Sumatra is part of Indonesia, but lies just across the Strait of Malacca from peninsular Malaysia (glossed below as Malaya). Important pre-colonial kingdoms in the region—including the Kingdom of Srivijaya and the Sultanate of Malacca—encompassed parts of both Sumatra and Malaya, and the Malay ethnic group today is one of the principal ethnic groups in several provinces in Sumatra. However, unlike in Malaysia, the Indonesian constitution affords no special rights to any ethnic

group. Comparing social beliefs about Malays in Malaya and Sumatra thus can provide novel insights into the ways in which different constitutional and political arrangements encourage different social beliefs about the same ethnic group.

Data for this exercise comes from 600 Indonesians from the provinces of Jambi, South Sumatra, Riau, and the city of Medan who surveyed in June 2017 (details about the survey are available in Appendix S1). Jambi, Riau, and South Sumatra were chosen because they are provinces where Malays have historically formed the majority population, while Medan was chosen because it is the closest large city. Respondents were asked identical questions about principal ethnic groups found in Sumatra as those in Malaysia. Responses were recorded in *Bahasa Indonesia*, the national language of Indonesia. Because Indonesian and Malay are two standardized versions of the same language, it is possible to analyze them at once.

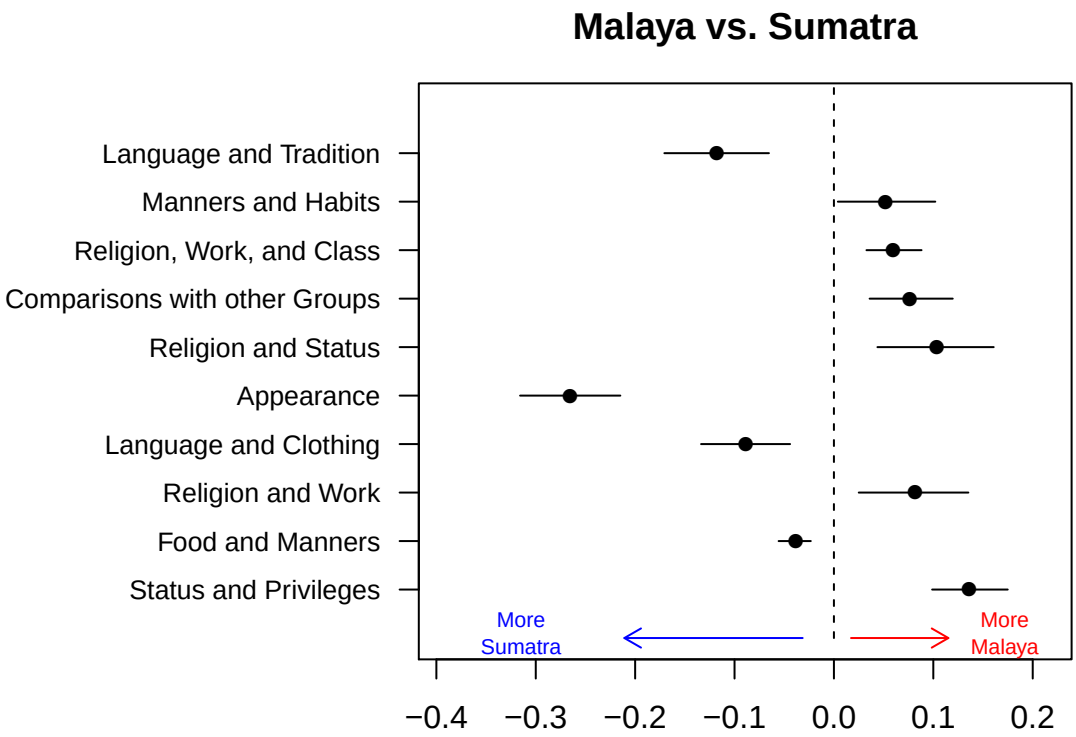
To prepare the data for processing, all responses for the non-Malay ethnic groups in the Malaysian and Indonesian datasets were discarded, as well as all responses from the Malaysian states of Sabah and Sarawak. I then proceed using the same topic modeling approach as described above, except for instead of exploring differences across ethnic groups, I focus on differences across national borders, captured with the variable $Region \in \{Malaya, Sumatra\}$, for one ethnic group:

$$\mathbf{B}_t = \boldsymbol{\alpha}_t + \sum_{j=1}^{J} \boldsymbol{\delta}_{jt} Demo_j + \boldsymbol{\gamma}_t Region + \sum_{j=1}^{J} \boldsymbol{\beta}_{jt} (Region \cdot Demo_j) + \mathbf{e}_t \tag{4}$$

Similar to Expression 3, Expression 4 provides us with $\boldsymbol{\gamma}_t$, a vector of parameters which measure the strength of the association between *Region* and belief t, and $\boldsymbol{\beta}_{jt}$, $T \times J$ matrix of parameters that capture differences in social beliefs about Malays across demographic groups j. As before, this procedure allows me to associate topics with regions; to answer the question of "what social beliefs about Malays are typically associated with peninsular Malaysia versus Sumatra?"

Figure 5 plots differences in predicted topic prevalence for each of the ten topics for respondents in Malaya versus Sumatra. The findings allow us to characterize important differences in

Figure 5: Cross-National Differences



Note: This figure plots the predicted difference in topic prevalence (and its 95% confidence interval) for Malays according to whether the respondent is in peninsular Malaysia or Sumatra.

the ways in which Indonesians in Sumatra and Malaysian in Malaya understand what it means to be a Malay. Beliefs about religion—specifically, Islam—are more likely to be found among Malaysian respondents than Sumatran respondents, even though nearly all Malays in Sumatra are Muslims too. Malaysians are also more likely to invoke notions of status or comparison with other ethnic groups than are Indonesians. By contrast, Indonesians in Sumatra are more likely to associate Malayness with language, physical appearance and dress, and food. These findings are consistent with a broader argument that Malaysia's constitution (which defines Malayness with respect to Islam) and political and economic order (which reserve special rights for Malays) have concrete effects on how ordinary people understand what it means to be Malay.

6 Conclusion

This paper has introduced a new approach for discovering social beliefs about ethnic structure and exploring heterogeneity in those beliefs. Building on new methodological tools for treating text as data, it enables fast discovery of socially meaningful beliefs about ethnic structure from short open-ended survey responses that are easily collected using standard survey techniques. Representing social beliefs about ethnic structure as a function of individual, demographic, and common societal beliefs, this approach provides a structural model of survey responses that facilitates a clear distinction of beliefs that are truly common across society and those that depend on the characteristics of survey respondents. Heterogeneity in social beliefs across ethnic groups and respondents is a natural feature of this model, and can be treated as a hypothesis to be tested in the data.

There are many directions to take this research. One is to look comparative across countries, to study how societies with a common inventory of ethnic identities end up with different social beliefs about ethnic structure. This will allow researchers to develop and test more refined hypotheses about ethnic politics than can be derived from group size and distribution alone. Another is exploit cross-cuttingness in ethnic, religious, and other identities (Selway 2011). In the Malaysian context it is relatively common for individuals to identify situationally as Malay or Arab, but much less common to identify situationally as Malay or Chinese. The tools introduced in this paper allow us to unpack the common social meanings that may be invoked for such cross-cutting ethnic identities as Malay and Arab, embracing the complexity of social identities without sacrificing the generality and falsifiability of a quantitative, survey-based approach to identity.

References

Adriani, Mirna et al. (2007). "Stemming Indonesian: A Confix-Stripping Approach". *ACM Transactions on Asian Language Information Processing* 6 (4), pp. 1–33.

Alatas, Syed Hussin (1977). *The Myth of the Lazy Native: A Study of the Image of the Malays, Filipinos and Javanese from the 16th to the 20th Century and Its Function in the Ideology of Colonial Capitalism*. London: Frank Cass.

Blei, David M. (2012). "Probabilistic Topic Models". *Communications of the ACM* 55 (4), 77?84.

Brubaker, Rogers et al. (2006). *Nationalist Politics and Everyday Ethnicity in a Transylvanian Town*. Princeton: Princeton University Press.

Chandra, Kanchan (2012). "Introduction". In: *Constructivist Theories of Ethnic Politics*. Ed. by Kanchan Chandra. New York: Cornell University Press, pp. 1–47.

Collins, James T. (1998). *Malay, World Language: A Short History*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Department of Statistics Malaysia (2016). "Current Population Estimates, Malaysia, 2014 - 2016". Available online at http://bit.ly/2jQiguN.

Esteban, Joan, Laura Mayoral, and Debraj Ray (2012). "Ethnicity and Conflict: Theory and Facts". *Science* 336, pp. 858–865.

Fearon, James D. (2003). "Ethnic and Cultural Diversity by Country". *Journal of Economic Growth* 8.2, pp. 195–222.

"Federal Constitution of Malaysia". Available online at http://bit.ly/1Wl4PSh.

Goldstein, Judith and Robert O. Keohane (1993). "Ideas and Foreign Policy: An Analytical Framework". In: *Ideas and Foreign Policy: Beliefs, Institutions, and Political Change*. Ed. by Judith Goldstein and Robert O. Keohane. Ithaca: Cornell University Press, pp. 3–30.

Mohamad, Maznah (2003). "Malaysia in 2002: Bracing for a Post-Mahathir Future". *Southeast Asian Affairs* 2003, 149?167.

Nagata, Judith A. (1974). "What Is a Malay? Situational Selection of Ethnic Identity in a Plural Society". *American Ethnologist* 1 (2), pp. 331–350.

Nazief, B.A.A. and M. Adriani (1996). *Confix-stripping: Approach to stemming algorithm for Bahasa Indonesia*. Tech. rep. Faculty of Computer Science, University of Indonesia.

Pepinsky, Thomas B., R. William Liddle, and Saiful Mujani (2018). *Piety and Public Opinion: Understanding Indonesian Islam*. New York: Oxford University Press.

Posner, Daniel N. (2004). "The Political Salience of Cultural Difference: Why Chewas and Tumbukas Are Allies in Zambia and Adversaries in Malawi". *American Political Science Review* 98.4, pp. 529–545.

Roberts, Margaret E., Brandon M. Stewart, and Edoardo M. Airoldi (2016). "A Model of Text for Experimentation in the Social Sciences". *Journal of the American Statistical Assocation* 111.515, pp. 988–1003.

Roberts, Margaret E. et al. (2014). "Structural Topic Models for Open-Ended Survey Responses". *American Journal of Political Science* 58.4, pp. 1064–1082.

Selway, Joel Sawat (2011). "The Measurement of Cross-cutting Cleavages and Other Multidimensional Cleavage Structures". *Political Analysis* 19.1, pp. 48–65.

Setiabudi, Nur Andi (2017). *katadasaR: An R Package of Word Stemming for Bahasa Indonesia Using Nazief & Adriani's Algorithm*. Tech. rep.

Smith, Tom W. et al. (2016). "General Social Surveys, 1972-2016 [machine-readable data file]". Available online at gssdataexplorer.norc.org.

Taylor, C.L. and M.C. Hudson (1972). *World Handbook of Political and Social Indicators*. New Haven, CT: Yale University Press.

S1 Survey Details

S1.1 Malaysia

The Malaysian survey was implemented by the Merdeka Centre for Public Opinion Research in March 2017. The survey combined two modes of contacting respondents: phone and in-person. Of 1210 survey respondents, 212 were contacted in person, exclusively in the rural northern states of Kedah, Kelantan, and Terengganu where landlines remain uncommon. The sampling is proportional to state population and urban-rural divide. The average interview time is 26 minutes, with a minimum of 7 minutes and a maximum of 2 hours 49 minutes. The breakdown of survey respondents by ethnic identity and state appears in Table S1-1.

Table S1-1: Respondent Ethnicity by State/Federal Territory

State	Chinese	Indian	Malay	Muslim Bumiputera	Non-Muslim Bumiputera
Johor	60	6	77	0	0
Kedah	12	4	76	0	0
Kelantan	1	0	81	0	0
Melaka	15	4	23	0	0
Negeri Sembilan	15	8	27	0	0
Pahang	14	2	49	0	0
Perak	53	17	66	0	0
Perlis	1	0	10	0	0
Pulau Pinang	41	10	28	0	0
Sabah	14	0	0	45	35
Sarawak	26	0	0	33	36
Selangor	67	30	95	0	0
Terengganu	2	0	58	0	0
Kuala Lumpur (F.T.)	35	5	29	0	0

S1.2 Indonesia

The Indonesia survey was implemented by Lembaga Survei Indonesia in June 2017. All respondents were interviewed in person by trained enumerators. The breakdown of survey respondents by ethnic identity and province or city appears in Table S1-2.

Table S1-2: Respondent Ethnicity by Province/City

	Medan City	Jambi	Riau	South Sumatra
Malay	14	63	62	108
Batak	45	6	15	1
Chinese	2	2	1	3
Minangkabau	13	9	27	0
Other non-Sumatran	76	70	45	38

S2 Different Numbers of Topics

The following two tables present highest probability topics for topic models that set T = 10 (Table S2-1) and T = 30 (Table S2-2).

Table S2-1: Ten Topics

Topic	HighestProb	Malay	Chinese	Arab	Indian
1	rajin yang lebih ekonomi usaha duit didik	0.072	0.147	0.058	0.096
2	islam kuat hati boleh dengki daripada percaya	0.096	0.056	0.096	0.076
3	kerja malaysia cina makan malaysian kasar muslim	0.118	0.128	0.211	0.159
4	dengan dan ada satu maju tolong sama	0.108	0.083	0.065	0.106
5	bahasa kulit cakap budaya arab gelap tinggi	0.105	0.116	0.177	0.168
6	kaum malas kaya kurang diri sopan sendiri	0.133	0.108	0.106	0.095
7	niaga baik banyak pandai amah tipu hormat	0.072	0.15	0.067	0.085
8	suka agama mudah lupa tak hindu pegang	0.132	0.087	0.103	0.106
9	orang tidak untuk padu sikap sahaja bumiputera	0.088	0.06	0.076	0.058
10	layu dalam bangsa lain hidup tiada mahir	0.076	0.065	0.042	0.049

Table S2-2: Thirty Topics

Topic	HighestProb	Malay	Chinese	Arab	Indian
1	kasar sikap susah hidup cantik bijak hadap	0.014	0.028	0.058	0.037
2	cina pemikiran lalu pilih tolong-menolong tegas teruk	0.015	0.038	0.003	0.01
3	yang ada jahat kari kecil buruk baikada	0.034	0.025	0.018	0.056
4	arab bahasa india nasi hidung comel lemak	0.023	0.006	0.113	0.036
5	banyak budaya tinggi adat rupa masalah paras	0.045	0.033	0.046	0.04
6	diri sendiri penting padu utama tamak kedekut	0.011	0.052	0.021	0.011
7	layu bahasa cakap tak cara tutur ikut	0.091	0.051	0.034	0.047
8	kerja dengan lain gaul libat bebas asa	0.055	0.054	0.028	0.053
9	agama negara pegang buddha sangat kepada asal	0.045	0.027	0.052	0.017
10	baik hati tetap murah gelintir masih licik	0.034	0.026	0.01	0.036
11	kurang maju usaha mesra lebih pengaruh lemah	0.049	0.032	0.025	0.039
12	kaum malaysia satu duit kawan majoriti ramai	0.042	0.051	0.026	0.044
13	orang lebih lain buka bukan daripada biasa	0.021	0.016	0.039	0.011
14	untuk mereka raja buat dapat senang sahaja	0.039	0.037	0.02	0.038
15	kaya tiada miskin ansur tolak manusia dua	0.01	0.024	0.053	0.028
16	niaga kuat pandai ajar cari bisnes kedai	0.021	0.104	0.055	0.047
17	dengki asing rendah etnik dengan kumpul semua	0.034	0.013	0.029	0.008
18	malas sopan santun amah jual karpet adab	0.09	0.012	0.044	0.023
19	rajin tipu putih ahli sepakat minda kerjasama	0.021	0.073	0.017	0.025
20	dalam bangsa mahir baju tanah aktif jenayah	0.025	0.029	0.019	0.016
21	muslim pakai berbeza sombong tudung putih masak	0.034	0.024	0.061	0.015
22	tidak mudah boleh lupa bantu percaya beri	0.054	0.035	0.029	0.046
23	islam bumiputera cepat datang konservatif marah pedas	0.046	0.005	0.044	0.002
24	suka tolong sosial bagus semangat sifat hibur	0.041	0.052	0.038	0.045
25	sama mata masa sepet halal roti ambil	0.021	0.033	0.014	0.035
26	makan malaysian budi hormat khinzir org syukur	0.031	0.02	0.009	0.02
27	cakap ganas gaduh belit gangster bising mabuk	0.011	0.011	0.035	0.073
28	kulit gelap besar hitam hindu warna tamil	0.015	0.025	0.04	0.115
29	ekonomi kuasa dari didik berdikari tinggal kukuh	0.02	0.055	0.012	0.02
30	dan kenal guam dlm curi cemburu matlamat	0.008	0.009	0.01	0.009

S3 Representative Responses

Table S3-1 lists five representative responses for the four topics discussed in the text.

Table S3-1: Five Representative Responses

Topic 1	English Gloss
tutur dalam bahasa arab	in Arabic
orang asing	foreigner
orang arab	Arab person
orang asing	foreigner
orang asing	foreigner

Topic 2	English Gloss
pegang kepada agama manusia tuah	depend on religion lucky person
pegang kuat pada agama ikut undang-undang islam penuh	hold on to religion follow Islamic law fully
gantung kepada raja ingin bebas dalam islam	depend on the government want to be free in Isl
bahagia raja sokong mereka	happy government supports them
selalu harap dapat subsidi raja malas	always hope to get government subsidies lazy

Topic 13	English Gloss
warna kulit gelap kuih maru	dark skin dhal snacks
warna kulit gelap cakap tamil	dark skin speak Tamil
cakap putar belit samseng (mabuk)	double talking drunk
deepavali sari	Deepavali sari
rupa paras gelap bahasa tamil	dark face Tamil language

Topic 16	English Gloss
rajin kerja ahli niaga	work hard business expert
niaga restoran	business restaurant
untung lampau	lucky past
ahli niaga	business expert
niaga kaya	business rich