Building Knowledge-Guided Lexica to Model Cultural Variation

Shreya Havaldar[†], **Salvatore Giorgi**[†], **Sunny Rai**[†], **Thomas Talhelm**[◇], **Sharath Chandra Guntuku**[†], **& Lyle Ungar**[†]
[†]University of Pennsylvania, [◇]University of Chicago
{shreyah,ungar}@seas.upenn.edu

Abstract

Cultural variation exists between nations (e.g., the United States vs. China), but also *within* regions (e.g., California vs. Texas, Los Angeles vs. San Francisco). Measuring this regional cultural variation can illuminate how and why people think and behave differently. Historically, it has been difficult to computationally model cultural variation due to a lack of training data and scalability constraints. In this work, we introduce a new research problem for the NLP community: *How do we measure variation in cultural constructs across regions using language?* We then provide a scalable solution: building knowledge-guided lexica to model cultural variation, encouraging future work at the intersection of NLP and cultural understanding. We also highlight modern LLMs' failure to measure cultural variation or generate culturally varied language.

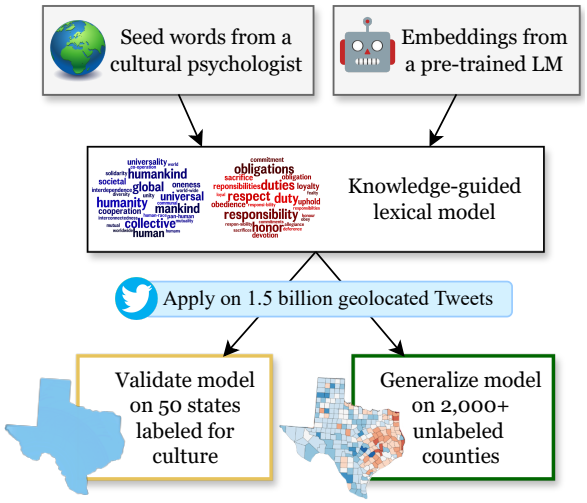

Figure 1: We build knowledge-guided lexica to model cultural variation. Our method encodes domain knowledge via seed words based on cultural psychology theory. We use embeddings to transform these seed words into a high-validity lexical model that successfully measures cultural variation across the US.

1 Introduction

People think and behave differently around the world. This is partly due to *cultural variation*, or the differences among individuals that exist due to some form of social learning (Cohen, 2001).

Having a computational method that utilizes language to measure cultural variation could help us better understand the way people communicate (Tsai et al., 2006; Oishi et al., 2009), build more culturally-aware NLP systems (Hovy and Yang, 2021), and advance interdisciplinary research in anthropology, cultural psychology, etc. However, due to a lack of data and scalability constraints, few such methods exist.

In this paper, we present *measuring regional variation in culture* as a problem of interest for the NLP community. We highlight how large language models (LLMs) struggle with this task, and build a knowledge-guided lexical model as a scalable and reliable solution. Specifically, we focus on measuring the cultural dimenision of *individualism and collectivism*[1] across the United States (US) using geolocated Tweets.

Historically, measuring cultural dimensions across regions has been mostly done through questionnaires, such as the World Values Survey (WVS) (Haerpfer et al., 2020). However, questionnaires are time-consuming and heavily restricted in scope; the most recent WVS wave required four years and averaged only 52 participants per US state. Recent work probes LLMs for cultural values (Arora et al., 2023), but these LLMs do not reflect all cultures equally (Havaldar et al., 2023b). Therefore, relying on LLMs to measure culture is risky, as they may not generalize well to different populations.

[1]Cultural psychologists have quantified axes on which culture differs, also called *cultural dimensions*. A key cultural dimension is called *individualism vs. collectivism* (Hofstede, 2011). Collectivism stresses the importance of the community, while individualism focuses on each person's rights and concerns. This dimension has been shown to influence behaviors like voting, donating, etc.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 211–226
June 16-21, 2024

The overhead of traditional survey-based approaches and inconsistent cultural awareness of existing LLMs motivate computational methods that rely on *existing language data* to measure cultural variation instead. We specifically seek to use geolocated Twitter data — instead of selecting a small portion of people to represent a state or county (like traditional survey-based methods), we instead use a massive amount of Tweets from that region, thus gaining a larger and more holistic representation of a region's population.

We also seek to leverage *domain expertise* from cultural psychologists. Cultural psychologists have spent decades developing non-computational tools to measure cultural constructs like individualism and collectivism (Talhelm et al., 2014). By encoding this domain knowledge via a set of expert-curated seed words, we can create a method to measure culture that is both scalable and grounded in cultural psychology theory.

In this paper, as an example of cultural variation, we will measure individualism and collectivism across US counties using the following resources:

- Domain knowledge from an expert psychologist researching collectivism.
- An open-source corpus (see Appendix A) of 1.5 billion geolocated Tweets from 6 million US users (Giorgi et al., 2018).
- Collectivism indicators (survey data, living arrangements, religiosity, and ingroup bias) to validate our results for fifty US states. (Vandello and Cohen, 1999; Pelham et al., 2022).

Challenges with deep learning approaches. A modern NLP solution to measure culture today would take the form of either labeling data and training a model, or prompting a pre-trained LM. However, LLMs have been shown to lack cultural awareness (Havaldar et al., 2023b; Liu et al., 2023), so cultural insights from these models may be incorrect or untrustworthy.

Additionally, classifying 1.5 billion Tweets requires a sizable amount of labeled training data, and training on such a large dataset is not computationally scalable. For instance, running our entire corpus through GPT-4 would cost roughly $900,000 (see Appendix B).

At a higher level, building a Tweet-level deep learning model to predict culture is impractical. Most of an individual's language does not indicate their cultural beliefs. Given this sparsity, labeling enough Tweets to train an adequate model is prohibitively expensive.

Our method builds upon a line of work in NLP called lexicon induction (Araque et al., 2020; Buechel et al., 2020; Geng et al., 2022; Havaldar et al., 2023a), which analyzes massive corpora in NLP without solely relying on deep learning. Past work mainly builds lexica for sentiment, emotion, etc. We uniquely focus on lexicon induction in the domain where little labeled data exists and not every utterance can be relevantly labeled.

Our contributions are as follows:

1. We present *measuring regional variation in culture* as a problem of interest for the NLP community.
2. We develop knowledge-guided lexical models and demonstrate their ability to measure individualism and collectivism. Our method (1) is highly scalable, (2) encodes domain knowledge from cultural psychology, and (3) does not require additional labeled data.
3. We validate our method against past collectivism research at the US state-level and present novel results at the US county level.
4. We provide new insights into cultural variation across the US via a taxonomy of *communities* (socio-demographic clusters of counties) from the American Communities Project (Chinni and Gimpel, 2011).
5. We highlight the failure of modern LLMs (GPT-3.5, GPT-4) to measure cultural variation or generate language that matches real-world cultural variation.

2 Building Knowledge-Guided Lexica

Issues with traditional lexica. Lexica, or sets of curated words, are a highly scalable and explainable method for analyzing large datasets. However, building a full lexicon linked to cultural theory from the ground up is a time-intensive process, and sometimes takes psychologists years to complete (Pennebaker et al., 2015). Additionally, psychologists often make the error of including words in lexica that correlate negatively with the construct they are trying to measure. Jaidka et al. (2020) find that removing certain words from lexical categories in LIWC 2015 (Pennebaker et al., 2015) actually improves overall performance.

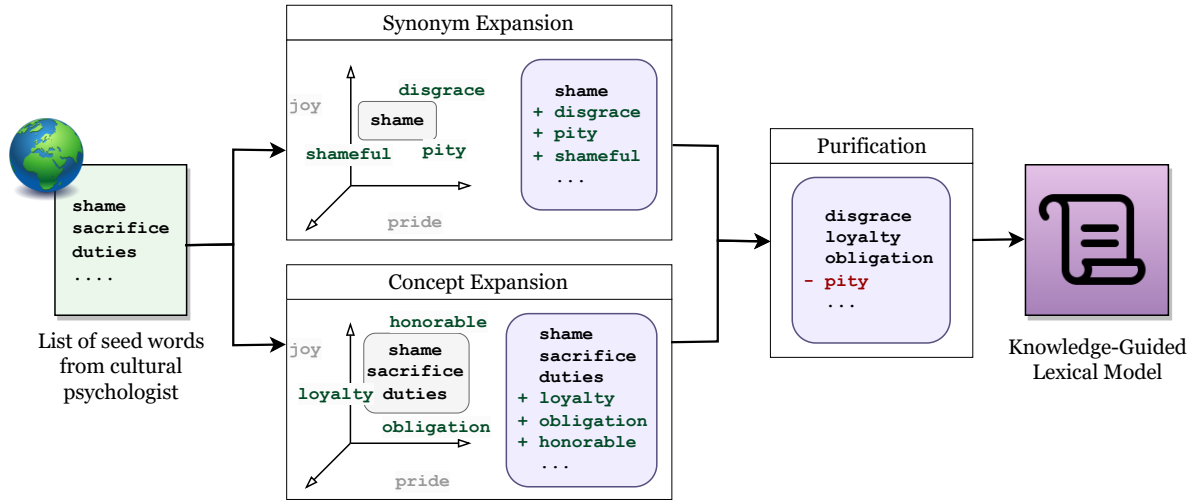

Figure 2: Our knowledge-guided lexica creation method. We begin with a set of seed words curated by an expert psychologist. The first stage, *Expansion*, consists of synonym expansion and concept expansion, done in parallel. The second stage, *Purification*, includes frequency-based and correlation-based pruning, done sequentially.

These erroneous words can be caused by the following two phenomena:

1. *Inflated Frequency:* Highly frequent words in a lexical category can correlate negatively with their counterparts (e.g. words like "love" and "lol", contained in the LabMT positive sentiment lexicon (Dodds et al., 2011) correlate negatively with happiness (Jaidka et al., 2020)) and muddy the results of a lexicon.
2. *Polysemy*: Words like "tender," which describe positive emotion, may have other meanings (e.g. describing chicken/steak, or referring to financial tenders), thus reducing the effectiveness of a lexicon.

To mitigate this, we propose a method that still builds on domain knowledge, but produces a lexicon that is *internally coherent*, bypassing the issues introduced by inflated frequency and polysemy.

Our approach has two components: Expansion and Purification. Figure 2 details this approach for the collectivism lexicon we generate. Using this method, we utilize domain knowledge from an expert psychologist to create a lexical model that has wider coverage of our corpus. We can then use this model to analyze our geolocated Twitter corpus and measure regional variation in individualism and collectivism across the US.

Step 1: Seed Word Generation We first ask an expert psychologist who has researched individualism and collectivism for many years to generate two small sets of seed words that capture each of these constructs (see Table 1). However, a small set of seed words may not be enough to sufficiently analyze a corpus of 1.5+ billion Tweets and may have erroneous words as described above.

Step 2: Expansion Next, we utilize word embeddings[2] to expand the set of seed words in two ways: we locate words that are similar to each seed word (*synonym expansion*), and locate the words that are similar to the overall construct described by the complete set of seed words (*concept expansion*).

For synonym expansion, we find the nearest neighbors for each seed word in embedding space and add these neighboring words to our lexica. For example, in Figure 2, we expand the word "shame" and add "disgrace", "shameful", and "pity" to our collectivism lexicon. We use cosine similarity to determine the nearest neighbors.

For concept expansion, we average the embeddings of each seed word set to find the *centroid embeddings*. For example, to find the collectivism centroid embedding, we would average the embeddings of "shame", "sacrifice", "duties", etc. Then we find the nearest neighbors of each centroid embedding. Concept expansion adds other words to a lexical category that are similar to the overall concept described by the seed word set.

By using embedding space to expand our lexica, we can additionally calculate a *weight* for each ex-

[2]We use FastText (Bojanowski et al., 2017) due to its fixed vocabulary size, efficient nearest neighbors functionality, and ability to find synonyms in context-free scenarios, but our methods are more general and agnostic to embedding type.

Collectivism Seed Words	`duties, responsibilities, role, fit in, community, sacrifice, shame, required, rules, honor, support, rely, loyal, respect, obedience`
Individualism Seed Words	`humans, humanity, worldwide, universal, mankind, everyone, collective, global, equity, imagination, cooperate, cooperation, shared, joint, identity, guilt, diversity`

Table 1: Seed words hypothesized to identify individualism and collectivism on social media, provided by an expert cultural psychologist. These words provide interesting insights into these two constructs — according to our domain expert, "collective" and "cooperation" are actually individualist words. This is because collectivism emphasizes close relationships over strangers, whereas individualism emphasizes strangers and weak ties, hence the usage of words like "collective".

panded word. The weight for each seed word is 1, and the weight for each word added during expansion is the cosine similarity to the corresponding seed word or centroid embedding. Highly similar words have a weight closer to 1, while more distant words have a lower weight. Our final lexicon is the union of seed words, words added via synonym expansion, and words added via concept expansion.

This method is highly tunable — any embeddings can be used, and the number of nearest neighbors returned during expansion can be adjusted based on the desired length of the final lexicon. To control the length of our final lexicon, we set two thresholds in this process: one for synonym expansion and one for concept expansion. We explore the effect of these expansion thresholds in Section 5.

Step 3: Purification Upon aggregating the words returned from both expansion types, we want to ensure that the resulting lexicon is both pertinent and internally correlated. Namely, we want to avoid the pitfalls of traditional lexica, where erroneous words may lower the overall performance.

To ensure pertinence, we filter out rare words, or any words below a given usage frequency (Bojanowski et al., 2017). Next, we ensure internal correlation. We apply our lexica to our US Twitter Corpus and compute the weighted frequencies for each word at the county-level.

Equation 1 details how we compute $F(w_i)$, the weighted frequency for a word w in County i, where T_i refers to the subset of Tweets geolocated in County i, and $\text{count}(w, t)$ refers to the number of times word w appears in Tweet t.

$$F(w_i) = \sum_{t \in T_i} \text{weight}_w * \text{count}(w, t) \quad (1)$$

To avoid issues that arise with inflated frequency and polysemy, we want to ensure that there are no words that correlate negatively with the other words in the lexicon. Specifically, we ensure that the product-moment correlation detailed in Equation 2 is greater than some positive threshold for every word w_i in our lexicon L.

$$r(F(w_i), \sum_{w \in L \setminus w_i} F(w)) \quad (2)$$

If a word does not meet this criteria, we remove it from the lexicon. This purification step ensures that every word contributes correctly to measuring the relevant cultural dimension. We explore the effect of this purification threshold in Section 5.

Figure 6 visualizes our expanded and purified knowledge-guided individualism and collectivism lexica. Note that this method can be used to measure regional variation for any cultural construct[3] by changing the seed words accordingly.

3 Evaluation

Upon expanding and purifying the lexica, we apply it to our Twitter Corpus. To get a collectivism score for county C, we sum the weighted frequencies of each word w in our collectivism lexicon L_{coll}, as outlined in Equation 3.

$$\text{Collectivism}(C) = \sum_{w \in L_{\text{coll}}} F(w) \quad (3)$$

We then aggregate these county-level scores to the state-level, and validate our results using past state-level collectivism research from Vandello and Cohen (1999); Pelham et al. (2022). We expect our state-level collectivism results to correlate positively with these indicators. As there are no similar resources for individualism, we use the strength of negative correlation to evaluate our state-level individualism results.

[3]Other example cultural constructs include power distance (Hofstede, 2011) or looseness/tightness (Gelfand et al., 2006) and are also hypothesized to vary regionally.

	Vandello & Cohen's Collectivism Scores	Grandparents *(GCI)*	Religiosity *(GCI)*	Ingroup Bias *(GCI)*	Average Validity
Collectivism (↑ is better)					
KGL Score (ours)	0.380*	0.362*	0.410*	0.467*	**0.405**
Seed Words Only	-0.033	-0.267	-0.352*	-0.142	-0.198
GPT-3.5 Baseline	-0.094	-0.094	0.056	-0.035	-0.042
Individualism (↓ is better)					
KGL Score (ours)	-0.379*	-0.571*	-0.659*	-0.515*	**-0.531**
Seed Words Only	-0.509*	-0.423*	-0.564*	-0.525*	-0.506
GPT-3.5 Baseline	-0.346*	-0.222	0.058	0.058	-0.113

Table 2: Pairwise product-moment correlations between our knowledge-guided lexica (KGL) scores and collectivism validation variables. We use Vandello & Cohen's Collectivism Scores and collectivism indicators from the Global Collectivism Index (GCI) at the US-state level. * indicates correlation is significant ($p < 0.05$). Underlined correlations are not found to be significantly different after bootstrapping test. We see that our method (KGL) outperforms both baselines — using only the seed words provided by an expert psychologist, and zero-shot labeling a subset of Tweets from each state using GPT-3.5.

Vandello & Cohen's Scores. Vandello and Cohen (1999) conduct a survey-based study of individualism and collectivism in the US and rank all states from most to least collectivist. We use these rankings as our first validation variable.

GCI Indicators. We also use relevant collectivism indicators from the Global Collectivism Index (Pelham et al., 2022) using corresponding questions from the 2017 US census and the 2017 wave of the World Values Survey (Haerpfer et al., 2020).

All six variables in the Global Collectivism Index – total fertility rate, living arrangements (% households with people over 60 and children under 14), stability of marriage (divorce rate to marriage rate ratio), religiosity, collective transportation, and ingroup bias (approximated by compatriotism due to lack of state-level data) – are replicable at the state-level using US census data and WVS data. Note that when aggregating US census data from county-level to state-level, we weight each county equally, due to disproportionate amounts of data coming from big cities.

In order to determine which of these six replicated variables also measures collectivism within the United States, we sample subsets of the six variables and use Cronbach's alpha to maximize internal consistency. We limit the subsets to size three or larger, following Pelham et al. (2022)'s validation of three collectivism indicators per nation. Upon exploring all possible subsets of size three or more, we find that the set of living arrangements, religiosity, and ingroup bias yields the highest Cronbach's alpha (0.702); we choose these as our additional validation variables.

Income. As a sanity check, we additionally validate against median household income at the state-level. Prior research has found that higher income levels lead to more individualistic values (Pelham et al., 2022). Similarly, we find that median household income correlates positively with our individualism lexicon scores (0.431) and negatively with our collectivism lexicon scores (-0.288).

4 Results

Table 2 contains the correlations between our collectivism and individualism lexicon scores and each of the four validation variables, across US states. We observe that collectivist word use positively correlates with all validation outcomes, and individualist word use correlates negatively. We also observe a strong negative correlation (-0.510) between our individualism and collectivism scores at the US state level.

To further assess the success of our method, we compare our correlations against the following baselines:

- **Seed Words Only:** To analyze the efficacy of expansion + purification, we compare against solely using the expert-curated seed words.
- **GPT-3.5 Baseline:** To explore whether our method outperforms prompting a pre-trained LM, we subsample a total of 100,000 Tweets (2,000 per state) from our corpus and have

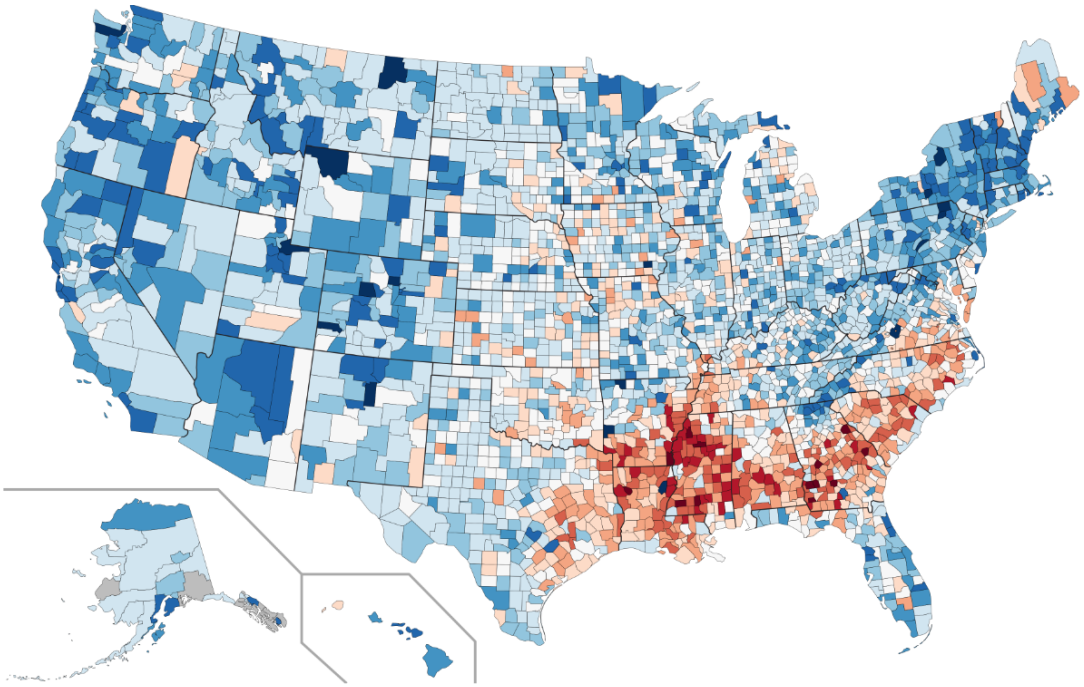

Figure 3: Collectivism (red) and individualism (blue) across US counties. Dark red = higher collectivism and dark blue = higher individualism. We include 2042 counties with sufficient data to compute individualism/collectivism scores, along with 1095 counties with interpolated scores based on geographic and socio-demographic variables.

GPT-3.5 label each Tweet as individualist, collectivist, or neither. We then calculate $\frac{num(\text{individualist})}{2{,}000}$ and $\frac{num(\text{collectivist})}{2{,}000}$ as the corresponding individualism and collectivism scores for that state. See Appendix C for additional details on prompting procedure.

Interpreting correlations in Table 2. There is little past work measuring culture in NLP, so we rely on state-level measures of collectivism to validate against, as no county-level measures exist. The magnitude of the effect sizes in Table 2 align with previous work. Giorgi et al. 2021 use this Twitter data to estimate 5-factor personality across US states. The average correlations across all five personality dimensions between these estimates and several national personality surveys range between 0.29 and 0.61.

Given our correlations fall within the range established in prior work, we conclude our lexica successfully capture individualism and collectivism.

Measuring county-level variation. Upon confirming the validity of our lexica, we apply them to county-level geolocated Tweets, as detailed in Equation 3, to gain a more fine-grained understanding of how individualism and collectivism vary regionally.[4] Figure 3 illustrates this variation, plotting the difference between the individualism and collectivism score. The deep south shows high levels of collectivism (dark red) and low levels of individualism (light blue). Conversely, the West Coast and the Northeast show low levels of collectivism (light red) and high levels of individualism (dark blue).

County interpolation. We interpolate individualism and collectivism across US counties where Twitter data is not available, in order to provide county-level estimates for the entire country.

To do this, we build a Gaussian Process (GP) regression model, which is traditionally used for spatial interpolation (i.e., kriging; Cressie, 1990). Instead of interpolating based on physical proximity alone, we follow Giorgi et al. 2023 and interpolate over both physical and socio-demographic space, training the GP model on latitude and longitude coordinates of the county centroids and 11 socio-demographic variables. See Appendix Section D for additional details.

Community-level insights. Cultural similarity is not always based on geographical proximity; two

[4]We release our lexica, county-level and state-level scores, and relevant code at https://github.com/shreyahavaldar/knowledge_guided_lexica

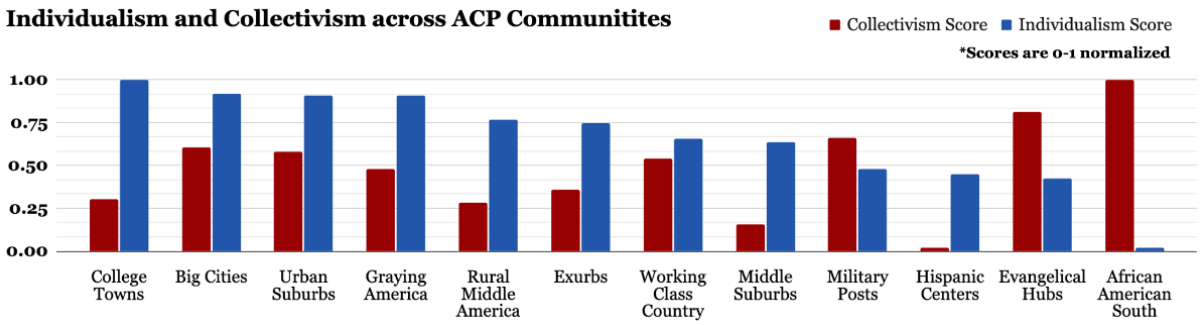

Figure 4: A comparison of collectivism (red) and individualism (blue) scores across communities defined by the American Communities Project, ordered from most individualistic (left) to least individualistic (right). We only analyze communities with over 40 included counties. *Scores are 0-1 normalized.*

cities hundreds of miles apart may be more similar than a city and a rural farm a few miles away (Guntuku et al., 2021). To show how county-level analyses of culture can help us better understand *communities*, we additionally use 15 community types (e.g., College Towns, Urban Suburbs) identified by the American Communities Project (ACP). The ACP identified these communities based on socio-demographic attributes, not spatial clusters of counties. Previous studies have used these community types to identify cultural variation in excessive alcohol consumption (Giorgi et al., 2020) and self-reported physical and mental health (Aggarwal et al., 2023; Mangalik et al., 2023).

Figure 4 shows county-level individualism and collectivism scores grouped into their corresponding ACP community (see Table A1 for counts.) These results provide novel insights into how culture varies regionally. For example, College Towns and Big Cities are highly individualist. These areas are also more affluent and have higher rates of education (ACP, 2023). This fits with prior research findings that people who are wealthy or educated tend to be more individualistic (Binder, 2019).

In contrast, the data shows that Evangelical Hubs and the African-American South are highly collectivist. These communities are tight-knit and religious areas (ACP, 2023), which have been linked to collectivism (Pelham et al., 2022). Military Posts are also more collectivist, which fits with the tight ties in military service and "duty to one's troop." This insight is helpful because we know of no cultural psychology research comparing military communities with civilian communities. Overall, our community-level findings are in line with prior work, and we introduce novel measurements for understudied communities.

5 Ablation Studies

5.1 Effect of Expansion Thresholds

To evaluate the effect of our expansion thresholds, we conduct a hyperparameter search with {0.7, 0.75, 0.8} as the search space for our synonym expansion threshold, and {0.4, 0.45, 0.5} as the sample space for our cluster expansion. We run our pipeline 9 times, testing each of these combinations. Table A2 contains the validation results for each of our hyperparameter runs.

We find that expansion greatly helps the validity of our collectivism lexicon, as our seed words do not perform well — a lower cosine similarity threshold yields higher validation scores. Interestingly, we observe the opposite phenomenon for our individualism seed words — more expansion hurts validity scores. We select {0.75, 0.45} for our final synonym and concept thresholds, respectively, as they yield the highest average validity.

As we see in the case of our collectivism lexicon, the expansion step is crucial to gathering the words that make our lexicon perform well. However, the results are unstable across thresholds. This is expected, as the resulting lexica are not internally coherent. Upon ensuring internal coherence, our lexica become significantly more stable.

5.2 Effect of Purification Threshold

To evaluate the effect of our purification threshold, we conduct another hyperparameter search over the space {0, 0.05, 0.1, 0.15, 0.2}. Here, we find that the purification threshold has very little impact on the overall validation score. As long as we ensure all words internally correlate positively (i.e. Equation 2 is greater than 0), the resulting lexicon is stable. We select 0.15 as our final purification threshold as this yields the highest average validity.

Table A3 contains the validation results for each purification threshold.

Some examples of words removed due to negative internal correlation include "benefit", "supporting", and "community" for collectivism, and "harmony", "evolution" and "international" for individualism. Though these words are added during expansion as synonyms of seed words, they have different usage patterns than their counterparts. Some are erroneous words due to polysemy — "benefit" may indicate helpful behavior, but may also refer to a fundraiser or gala event. Likewise, "harmony" may indicate unity, but is more frequently used to refer to music and melodies.

The purification component of our pipeline results in lexica that are highly stable across thresholds as well as reliable, highlighting the importance of internal coherence.

6 Investigating Cultural Variation in LLM-generated Text

With a validated method to measure regional variation in individualism and collectivism, we can now answer the question: *Given geographical information, can LLMs generate text that mimics real-world cultural variation?*

Generating state-specific text. A recent line of NLP research investigates how LLMs express personality, and explores imbuing a fixed set of characteristics into an LLM to create a persona (Mao et al., 2023; Safdari et al., 2023; Jiang et al., 2024). Drawing from this line of work, we create a geographic persona for an LLM (i.e. specifying a US state of residency), thus allowing us to generate state-specific text.

Specifically, we aim to recreate our dataset using an LLM, so we can directly compare whether synthetic LLM-generated text reflects the cultural variation found in Tweets from real people. We select four states for this experiment – New York, Massachusetts, Louisiana, and Mississippi, as they have the highest levels of individualism (NY, MA) and collectivism (LA, MS), while also containing Tweets from the vast majority of their counties.

Next, we prompt GPT-3.5 to generate Tweets as Twitter users living in each state. Following Jiang et al. (2024), we use a temperature of 0.7 to encourage creativity and variance when emulating different users. We keep our geographic persona prompt concise and open-ended, so as not to skew the LLM with prior notions of expected topics or writing style. We generate 100,000 total Tweets (500 users per state; 50 Tweets per user) and sample a parallel subset from our real-world Twitter dataset with an identical user breakdown.

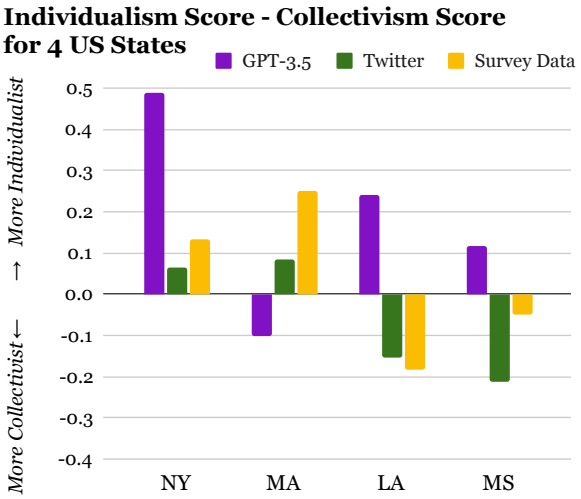

Figure 5: Individualism score minus collectivism score for LLM-generated and real-world Tweets. Across four US states, Twitter data (green) more closely aligns with Vandello & Cohen's survey-based scores (yellow) compared to the GPT-3.5 data (purple).

See Appendix E for full details on experimental setup and prompting procedure. Sample Tweets generated by GPT-3.5 are shown below:

```
Nothing beats a slice of New York pizza
on a Friday night #NYCeats
```

```
Louisiana summers are no joke, y'all.
The humidity is on another level.  Stay
cool out there, friends!  #LouisianaLife
```

LLMs don't reflect cultural variation. We then run our knowledge-guided lexical model on both datasets, and calculate the difference between the individualism and collectivism scores per state. The results are shown in Figure 5, along with Vandello & Cohen's collectivism scores[5].

The Twitter scores match what we expect: NY and MA are more individualist, while LA and MS are more collectivist. However, the GPT scores do not reflect this pattern. Upon inspecting the generated Tweets, we notice a deeper problem – the Tweets predominantly focus on state stereotypes (e.g. NY Tweets referencing pizza and bagels, LA Tweets referencing crawfish and Mardi Gras, etc.)

In fact, this phenomenon causes New York to have a very high individualism score, due to fre-

[5]As Vandello & Cohen's scores range from $0-1$, we shift the range by subtracting them from 0.5, thus ensuring a higher number indicates more individualism.

quently appearing phrases like "the greatest city in the world" and "diversity of cultures." However, the unstable and incorrect results for the other three states indicate that GPT-3.5 cannot reliably mimic real-world cultural variation. Rather, the generated Tweets are highly stereotypical and cover a small fraction of topics found in Tweets from real people.

7 Related Work

Lexicon Induction. Developing lexica to measure psychological and social constructs is a computationally inexpensive venture that can provide results at par with sophisticated LLMs. Emotion lexica (Mohammad and Turney, 2010), Demographic lexica (Sap et al., 2014), and Politeness lexica (Hayati et al., 2021; Havaldar et al., 2023a) are a few such examples. Similar to prior work, we begin with an expert-curated list of seed words that we expand using semantic knowledge in LLMs.

Measuring Cultural Constructs. Individualism and collectivism have previously been studied to understand a range of constructs (Hamilton et al., 2016; Hofstede, 2011; Triandis, 1993).

One of the earliest attempts to measure collectivism in the US uses a questionnaire-based survey approach (Vandello and Cohen, 1999). More recently, Bazzi et al. (2020) uses infrequent names (common names indicate the desire to fit in vs. stand out (Twenge et al., 2010)) as an indicator of individualism in their county-level study. Chen et al. (2021) use citizen ancestry data to account for immigrants' culture. However, static approaches including name and ancestry mapping ignore the transient nature of the collectivism-individualism dimension and the fact that it evolves with the socio-economic environment (Lomas et al., 2023; Santos et al., 2017). Question-based surveys may provide a truer picture but can become increasingly expensive and time-consuming when collecting granular data (e.g. county-level). In this scenario, social media language can help in dynamically modeling collectivism within regions (Aggarwal et al., 2023).

Culture and NLP. Detecting culture in LLMs and the consequent goal of "cultural alignment" are emerging research problems that often rely on decade-old measures of cultural constructs derived from a small sample of the population (Kovač et al., 2023; Jin et al., 2023; Masoud et al., 2023). Social psychology has great potential in adapting machine-generated text for cross-cultural interactions (Marmolejo-Ramos and Tejada, 2023). Through this paper, we propose a highly scalable approach to dynamically measure cultural constructs at high granularity from publicly available social media language.

8 Conclusion & Future Work

We introduce a new problem for the NLP community: *measuring regional variation in culture*. This is a new class of problem — obtaining labels for culture across states or countries is often infeasible, as there are very few labeled data points to train on. Using pre-trained LLMs to label data is unreliable, as these models lack basic cultural awareness. Additionally, scaling to label billions of data points is computationally infeasible. Classic deep learning methods fail to solve this problem; therefore, measuring culture necessitates a different approach.

We present a method to efficiently measure cultural variation by leveraging domain knowledge from cultural psychology to create knowledge-guided lexica. Our lexica filter out erroneous words and ensure internal coherence, bypassing the pitfalls of traditional lexica. By applying these lexica to social media language, we can estimate cultural differences at fine-grained geographic levels, such as states, counties, and communities – a task that modern LLMs fail to accomplish.

Future work could build on this method to get deeper insights into communities and cultures. For example, our method could be used to identify Tweets that mark cultural differences; we encourage researchers to build more sophisticated models on these identified Tweets. Additionally, our method is easily extendable to other cultural dimensions, such as power distance, tightness/looseness, etc. This method could also measure cultural variation globally, which requires analyzing different languages. Since our method is language agnostic, it can easily extend to non-English settings by leveraging multilingual embeddings.

9 Limitations

While we label each county for individualism and collectivism, we note that regions do not have a single culture. Within all regions, there is heterogeneity of cultural values and beliefs. Since we use an open-source Twitter corpus, we also have poor coverage of counties with little to no Twitter data. We also only use Twitter users to represent each region, which may lead to an incomplete represen-

tation of these regions. Additionally, not all aspects of culture are revealed in language – we are limited to analyzing only what people say online.

For our seed words, we only consult one domain expert. As a result, our final lexica are based on this expert's interpretation of individualism and collectivism. Extensions of this work could consult multiple experts to ensure downstream lexica are as unbiased as possible.

Furthermore, our results are influenced by the choice of embedding model used for expansion. Additional work is needed to determine the effect of the embedding model, namely, whether a different model yields drastically different results. We also risk propagating any bias present in the embedding model during expansion.

In our analyses, we do not control for race, income, or other demographic variables. We know cultural values are correlated with some demographic variables — for example, collectivism and individualism vary with income. Future work can improve upon these estimates by accounting for individual demographics. Additionally, it is unclear if this method of measuring cultural variation will work for all cultural dimensions. For example, power distance (Hofstede, 2011) involves the relationship dynamics of two people, which might make it difficult to capture with lexica.

10 Ethical Considerations

The goal of studying cultural variation is to better understand cultures, not individuals. Nonetheless, the characterization of culture has the danger of stereotyping individuals. Individuals within each culture vary greatly. Studying culture can help us understand differences in psychology, but we should not assume that a cultural average will definitely apply to a particular individual from that culture.

All data used in this study is publicly available. While geolocated Twitter data is used, only aggregated spatial-level data is reported. That is, no person-level identifiable information is used or released for this study.

References

ACP. 2023. Home - american communities project - americancommunities.org. https://www.americancommunities.org/. [Accessed 14-08-2023].

Arnav Aggarwal, Sunny Rai, Salvatore Giorgi, Shreya Havaldar, Garrick Sherman, Juhi Mittal, and Sharath Chandra Guntuku. 2023. A cross-modal study of pain across communities in the united states. In *Companion Proceedings of the ACM Web Conference 2023*, pages 1050–1058.

Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2020. Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-based systems*, 191:105184.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Samuel Bazzi, Martin Fiszbein, and Mesay Gebresilasse. 2020. Frontier culture: The roots and persistence of "rugged individualism" in the united states. *Econometrica*, 88(6):2329–2368.

Carola Conces Binder. 2019. Redistribution and the individualism–collectivism dimension of culture. *Social Indicators Research*, 142(3):1175–1192.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Sven Buechel, Susanna Rücker, and Udo Hahn. 2020. Learning and evaluating emotion lexicons for 91 languages. *arXiv preprint arXiv:2005.05672*.

Chinchih Chen, Carl Benedikt Frey, and Giorgio Presidente. 2021. Culture and contagion: Individualism and compliance with covid-19 policy. *Journal of economic behavior & organization*, 190:191–200.

Dante Chinni and James Gimpel. 2011. *Our patchwork nation: The surprising truth about the" real" America*. Penguin.

Dov Cohen. 2001. Cultural variation: considerations and implications. *Psychological bulletin*, 127(4):451.

Noel Cressie. 1990. The origins of kriging. *Mathematical geology*, 22:239–252.

Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one*, 6(12):e26752.

Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. 2018. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. *Advances in neural information processing systems*, 31.

Michele J Gelfand, Lisa H Nishii, and Jana L Raver. 2006. On the nature and importance of cultural tightness-looseness. *Journal of applied psychology*, 91(6):1225.

Yilin Geng, Zetian Wu, Roshan Santhosh, Tejas Srivastava, Lyle Ungar, and João Sedoc. 2022. Inducing generalizable and interpretable lexica. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4430–4448.

Salvatore Giorgi, Johannes C. Eichstaedt, Daniel Preotiuc-Pietro, Jacob R. Gardner, H. Andrew Schwartz, and Lyle H. Ungar. 2023. Filling in the white space: Spatial interpolation with gaussian processes and social media data. *Current Research in Ecological and Social Psychology*, page 100159.

Salvatore Giorgi, Khoa Le Nguyen, Johannes C Eichstaedt, Margaret L Kern, David B Yaden, Michal Kosinski, Martin EP Seligman, Lyle H Ungar, H Andrew Schwartz, and Gregory Park. 2021. Regional personality assessment through social media language. *Journal of Personality*.

Salvatore Giorgi, Daniel Preoţiuc-Pietro, Anneke Buffone, Daniel Rieman, Lyle Ungar, and H. Andrew Schwartz. 2018. The remarkable benefit of user-level aggregation for lexical-based population-level predictions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1167–1172, Brussels, Belgium. Association for Computational Linguistics.

Salvatore Giorgi, David B Yaden, Johannes C Eichstaedt, Robert D Ashford, Anneke EK Buffone, H Andrew Schwartz, Lyle H Ungar, and Brenda Curtis. 2020. Cultural differences in tweeting about drinking across the us. *International journal of environmental research and public health*, 17(4):1125.

Sharath Chandra Guntuku, Alison M. Buttenheim, Garrick Sherman, and Raina M. Merchant. 2021. Twitter discourse reveals geographical and temporal variation in concerns about covid-19 vaccines in the united states. *Vaccine*, 39(30):4034–4038.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, Bi Puranen, et al. 2020. World values survey: Round seven–country-pooled datafile. *Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat*, 7:2021.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Austin, Texas. Association for Computational Linguistics.

Shreya Havaldar, Matthew Pressimone, Eric Wong, and Lyle Ungar. 2023a. Comparing styles across languages.

Shreya Havaldar, Sunny Rai, Bhumika Singhal, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023b. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

Shirley Anugrah Hayati, Dongyeop Kang, and Lyle Ungar. 2021. Does bert learn as humans perceive? understanding linguistic styles through lexica. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6323–6331.

Geert Hofstede. 2011. Dimensionalizing cultures: The hofstede model in context. *Online readings in psychology and culture*, 2(1):8.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Kokil Jaidka, Salvatore Giorgi, H Andrew Schwartz, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2020. Estimating geographic subjective well-being from twitter: A comparison of dictionary and data-driven language methods. *Proceedings of the National Academy of Sciences*, 117(19):10165–10171.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Jad Kabbara, and Deb Roy. 2024. Personallm: Investigating the ability of large language models to express personality traits.

Chuanyang Jin, Songyang Zhang, Tianmin Shu, and Zhihan Cui. 2023. The cultural psychology of large language models: Is chatgpt a holistic or analytic thinker? *arXiv preprint arXiv:2308.14242*.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2023. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings.

Tim Lomas, Pablo Diego-Rosell, Koichiro Shiba, Priscilla Standridge, Matthew T Lee, Brendan Case, Alden Yuanhong Lai, and Tyler J VanderWeele. 2023. Complexifying individualism versus collectivism and west versus east: Exploring global diversity in perspectives on self and other in the gallup world poll. *Journal of Cross-Cultural Psychology*, 54(1):61–89.

Siddharth Mangalik, Johannes C Eichstaedt, Salvatore Giorgi, Jihu Mun, Farhan Ahmed, Gilvir Gill, Adithya V Ganesan, Shashanka Subrahmanya, Nikita Soni, Sean AP Clouston, et al. 2023. Robust language-based mental health assessments in time and space through social media. *arXiv preprint arXiv:2302.12952*.

Shengyu Mao, Ningyu Zhang, Xiaohan Wang, Mengru Wang, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2023. Editing personality for llms. *arXiv preprint arXiv:2310.02168*.

Fernando Marmolejo-Ramos and Julian Tejada. 2023. Social psychology: Spotting social faux pas with ai. *Communications Psychology*, 1(1):17.

Reem I Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2023. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. *arXiv preprint arXiv:2309.12342*.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34.

Shigehiro Oishi, Ed Diener, Richard E Lucas, and Eunkook M Suh. 2009. Cross-cultural variations in predictors of life satisfaction: Perspectives from needs and values. *Culture and well-being: The collected works of Ed Diener*, pages 109–127.

Brett Pelham, Curtis Hardin, Damian Murray, Mitsuru Shimizu, and Joseph Vandello. 2022. A truly global, non-weird examination of collectivism: The global collectivism index (gci). *Current Research in Ecological and Social Psychology*, 3:100030.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Henri C Santos, Michael EW Varnum, and Igor Grossmann. 2017. Global increases in individualism. *Psychological science*, 28(9):1228–1239.

Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and H Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1146–1151.

Taylor Sorensen, Liwei Jiang, Jena Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. 2023. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties.

Thomas Talhelm, Xuemin Zhang, Shigehiro Oishi, Chen Shimin, Dongyuan Duan, Xuezhao Lan, and Shinobu Kitayama. 2014. Large-scale psychological differences within china explained by rice versus wheat agriculture. *Science*, 344(6184):603–608.

Harry C Triandis. 1993. Collectivism and individualism as cultural syndromes. *Cross-cultural research*, 27(3-4):155–180.

Jeanne L Tsai, Brian Knutson, and Helene H Fung. 2006. Cultural variation in affect valuation. *Journal of personality and social psychology*, 90(2):288.

Jean M Twenge, Emodish M Abebe, and W Keith Campbell. 2010. Fitting in or standing out: Trends in american parents' choices for children's names, 1880–2007. *Social Psychological and Personality Science*, 1(1):19–25.

Joseph A Vandello and Dov Cohen. 1999. Patterns of individualism and collectivism across the united states. *Journal of personality and social psychology*, 77(2):279.

A Open-Source Twitter Corpus

We use the County Tweet Lexical Bank, an open source data set of features extracted from a corpus of 1.5 billion tweets from approximately 6 million US county-mapped users (Giorgi et al., 2018). While the full details of the dataset can be found in the original paper, we give a high-level summary to aid the reader. The dataset is built from a larger corpus which is a 10% sample of Twitter from 2009-2015 (over 30 billion tweets). These tweets are then mapped to US counties via latitude and longitude coordinates associated with the tweets or self-reported location information in the Twitter user's profile (a free text field). A Twitter user is included in this data set if they have posted at least 30 or more English tweets, and a county is included if at least 100 such users are mapped to that respective county. This process resulted in 1.5 billion tweets mapped to over 2,000 US counties.

B Scalability Calculations

We outline the proposed costs of using various LM-based techniques to label our corpus of 1.5 billion Tweets:

Proposed cost of GPT-4 As of August 2023, the OpenAI API rate for GPT-4 is $0.06 cents per 1,000 tokens. Assuming 10 tokens per Tweet, we get:

$$1.5e9\,\text{Tweets} \times \frac{10\ \text{Tokens}}{\text{Tweet}} \times \frac{\$0.06}{1,000\ \text{Tokens}} \quad (4)$$

This yields a total cost of $900,000.

Proposed cost of GPT-3.5 As of August 2023, the OpenAI API rate for GPT-3.5 is $0.002 per 1,000 tokens. Assuming 10 tokens per Tweet, we get:

$$1.5e9\,\text{Tweets} \times \frac{10\ \text{Tokens}}{\text{Tweet}} \times \frac{\$0.002}{1,000\ \text{Tokens}} \quad (5)$$

This yields a total cost of $30,000.

C GPT-3.5 Baseline

To assess whether a pre-trained LLM is capable of measuring individualism and collectivism, we sample a subset of our Twitter corpus (2,000 Tweets per state) and have GPT-3.5 assign a label to each Tweet. Our prompt is as follows:

```
System Prompt: Given a Tweet, try to
reason about whether it reflects the
cultural dimension of Individualism
or Collectivism. Collectivists are
closely linked individuals who view
themselves primarily as parts of a
whole, be it a family, a network of
co-workers, a tribe, or a nation. Such
people are mainly motivated by the norms
and duties imposed by the collective
entity. Individualists are motivated by
their own preferences, needs, and rights,
giving priority to personal rather than
group goals.

If the Tweet does not reflect either
cultural dimension, please label it
'Neither'.

Example input and output:
Tweet: {Tweet Text}
Label: {Individualism, Collectivism,
Neither}

User Prompt: Tweet: {Tweet Text}
```

We provide GPT-3.5 with the definition of individualism and collectivism as defined by Triandis (1993), and then have it label each tweet with one of three labels – individualism, collectivism, or neither. Across all states, GPT-3.5 labels at least 50% of the Tweets, ensuring enough signal for adequate comparison. We take cues from Sorensen et al. (2023), which prompts LLMs for values, to design this prompt.

ACP Community	Num Counties
Exurbs	207
Graying America	164
African American South	252
Evangelical Hubs	269
Working Class Country	159
Military Posts	70
Urban Suburbs	103
College Towns	151
Big Cities	46
Hispanic Centers	87
Rural Middle America	403
Middle Suburbs	77

Table A1: Number of included counties for each ACP community included in the analysis in Figure 4.

D Interpolations

We use 11 socio-demographic variables to interpolate individualism and collectivism across US counties with insufficient Twitter data. This includes four socioeconomic variables (median household income, percentage of the population with a Bachelor's degree, unemployment rate, and high school graduation rate) and seven demographic variables: (population density, median age, and the percentage of the population in rural areas, Hispanic, fe-

Figure 6: Word clouds visualizing our individualism lexica (blue, top) and collectivism lexica (red, bottom). Larger words have a higher weight, while smaller words have a lower weight.

male, married, and African American). This is implemented with the GPyTorch package (Gardner et al., 2018) with a learning rate of 0.1 and 500 iterations over the training data.

Figure 7 shows the original 2042 counties, and Figure 3 contains the interpolated counties as well.

E GPT-3.5 Tweet Generation

To assess whether a pre-trained LLM can generate language that reflects real-world variation in individualism and collectivism, we have GPT-3.5 generate 100,000 Tweets from four states - New York, Massachusetts, Louisiana, and Mississippi.

Generating geographic personas. When constructing our prompt, we instruct the LLM to behave like an individual living in a given state. We specify solely the state of residency, keeping the prompt concise and open-ended, so as not to bias the LLM towards a specific topic or style when generating text.

Emulating users. To construct a dataset parallel to our Twitter dataset, we generate Tweets in batches, with each query to GPT-3.5 emulating a single "user". To encourage creativity and variance in responses, we set the temperature relatively high (0.7), following Jiang et al. (2024). Each query emulates a different "user," with 50 Tweets generated in response to a single query.

Our prompt is as follows:

```
System Prompt: You are an individual
who lives in {STATE}. You enjoy writing
Tweets and posting them on Twitter for
your followers to read.

User Prompt: Generate 50 Tweets
that you might write and post online.
Remember that each Tweet has a 280
character limit. Separate each Tweet
with a newline character.
```

We ensure that the resulting response meets the following criteria: (1) The response contains 50 separate Tweets, and (2) Each Tweet contains no more than 280 characters. All generated Tweets can be found at https://github.com/shreyahavaldar/knowledge_guided_lexica.

Cluster Expansion Threshold	Synonym Expansion Threshold	Lexicon Length	V&C's Collectivism Scores	Grandparents *(GCI)*	Religiosity *(GCI)*	Ingroup Bias *(GCI)*	Average Validity
Collectivism (↑ is better)							
0.4	0.7	366	0.273	0.504*	0.531*	0.552*	0.465
0.4	0.75	366	0.273	0.504*	0.531*	0.552*	0.465
0.4	0.8	366	0.273	0.504*	0.531*	0.552*	0.465
0.45	0.7	100	0.226	0.113	0.124	0.274	0.184
0.45	0.75	98	0.275	0.171	0.175	0.314*	0.234
0.45	0.8	95	0.263	0.168	0.167	0.305*	0.226
0.5	0.7	42	-0.014	-0.234	-0.297	-0.094	-0.16
0.5	0.75	39	0.029	-0.184	-0.26	-0.057	-0.118
0.5	0.8	34	0.008	-0.196	-0.279	-0.078	-0.137
Individualism (↓ is better)							
0.4	0.7	479	0.095	0.512*	0.446*	0.308*	0.34
0.4	0.75	479	0.095	0.512*	0.446*	0.308*	0.34
0.4	0.8	479	0.095	0.512*	0.446*	0.308*	0.34
0.45	0.7	119	-0.417*	-0.545*	-0.664*	-0.542*	-0.542
0.45	0.75	119	-0.417*	-0.545*	-0.664*	-0.542*	-0.542
0.45	0.8	119	-0.417*	-0.545*	-0.664*	-0.542*	-0.542
0.5	0.7	37	-0.545*	-0.504*	-0.618*	-0.539*	-0.552
0.5	0.75	37	-0.545*	-0.504*	-0.618*	-0.539*	-0.552
0.5	0.8	37	-0.545*	-0.504*	-0.618*	-0.539*	-0.552

Table A2: Ablation study investigating the effect of the expansion thresholds.

Purification Threshold	Lexicon Length	V&C's Collectivism Scores	Grandparents *(GCI)*	Religiosity *(GCI)*	Ingroup Bias *(GCI)*	Average Validity
Collectivism (↑ is better)						
0	62	0.385*	0.346*	0.397*	0.467*	0.399
0.05	56	0.391*	0.351*	0.404*	0.47*	0.404
0.1	43	0.388*	0.345*	0.4*	0.464*	0.399
0.15	30	0.38*	0.362*	0.41*	0.467*	0.405
0.2	18	0.365*	0.357*	0.412*	0.468*	0.4
Individualism (↓ is better)						
0	59	-0.35*	-0.554*	-0.656*	-0.512*	-0.518
0.05	56	-0.351*	-0.55*	-0.656*	-0.511*	-0.517
0.1	48	-0.374*	-0.568*	-0.658*	-0.513*	-0.528
0.15	42	-0.379*	-0.571*	-0.659*	-0.515*	-0.531
0.2	36	-0.382*	-0.569*	-0.66*	-0.52*	-0.533

Table A3: Ablation study investigating the effect of the purification threshold.

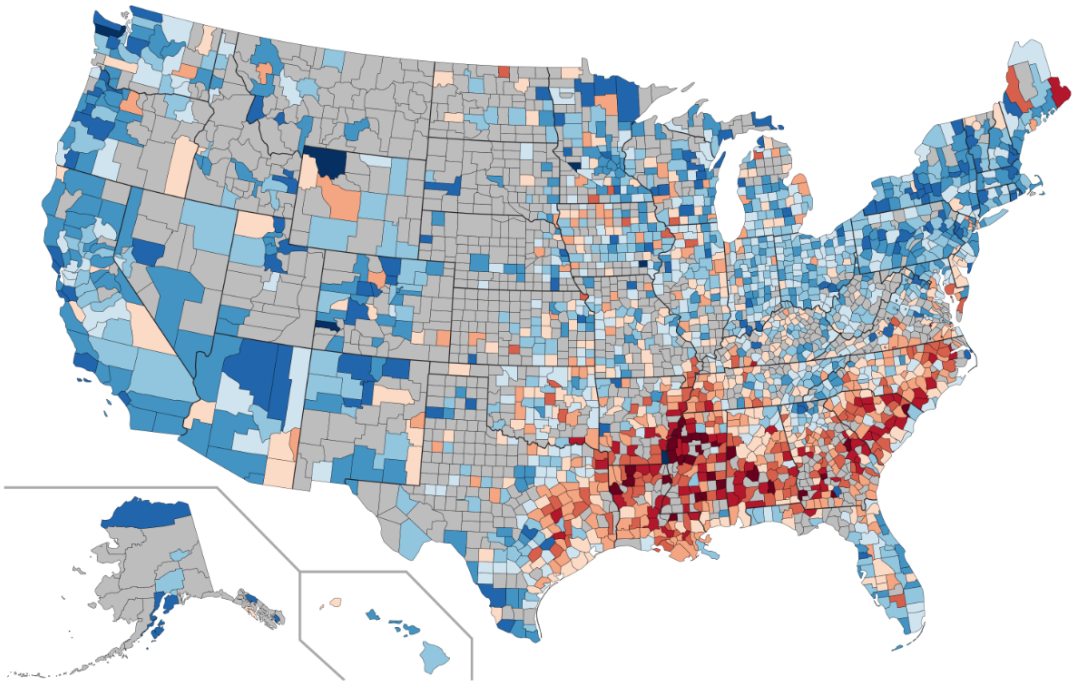

Figure 7: Collectivism (red) and individualism (blue) across US counties. Dark red = higher collectivism and dark blue = higher individualism. We show only the 2042 counties with sufficient data to compute individualism/collectivism scores. Gray counties do not have enough Twitter data to estimate scores.