America Tweets China: A Fine-Grained Analysis of the State and Individual Characteristics Regarding Attitudes towards China

Yu Wang Department of Political Science University of Rochester Rochester, NY, 14627, USA ywang176@ur.rochester.edu Jianbo Yuan Department of Computer Science University of Rochester Rochester, NY, 14627, USA jyuan10@ur.rochester.edu Jiebo Luo Department of Computer Science University of Rochester Rochester, NY, 14627, USA jluo@cs.rochester.edu

Abstract—The U.S.-China relationship is arguably the most important bilateral relationship in the 21st century. Typically it is measured through opinion polls, for example, by Gallup and Pew Institute. In this paper, we propose a new method to measure U.S.-China relations using data from Twitter, one of the most popular social networks. Compared with traditional opinion polls, our method has two distinctive advantages. First, our sample size is significantly larger. National opinion polls have at most a few thousand samples. Our data set has 724,146 samples. The large size of our data set enables us to perform state level analysis, which so far even large opinion polls have left unexplored. Second, our method can control for fixed state and date effects. We first demonstrate the existence of inter-state and inter-day variances and then control for these variances in our regression analysis. Empirically, our study is able to replicate the stylized results from opinion polls as well as generate new insights. At the state level, we find New York, Michigan, Indiana and Arizona are the top four most Chinafriendly states. Wyoming, South Dakota, Kansas and Nevada are most homogeneous. At the individual level, we find attitudes towards China improve as an individual's Twitter experience grows longer and more intense. We also find individuals of Chinese ethnicity are statistically more China-friendly.

Keywords- U.S.-China relations; Perceptions; Tweets; Sentiment Analysis;

I. INTRODUCTION

In international relations, perceptions matter [1], [2]. And how the U.S. perceives China matters particularly, the former being the world's only superpower and the latter a potential challenger. Typically these perceptions are measured using opinion polls. For example, a February 2015 survey by *Foreign Policy* shows that the majority of American students who studied in China have developed a more positive view of the country (sample size: 343) [3]. A February 2014 poll survey by *Gallup* shows that 53% of Americans view China very or mostly unfavorably (sample size: 1,023) [4].

While opinion polls have been the standard way for gauging public opinion, their weaknesses are obvious. First and foremost, the small sample size, as evidenced above, renders any fine grained analysis extremely difficult if not impossible [5]. Second, these opinion polls, mostly carried out on an annual basis, are susceptible to the influence of daily events. Surveys before or after the event can yield drastically different results, and yet both will be used as representing the annual result. In this paper, we propose a method that can solve these two problems. We measure U.S. perceptions of China through Twitter.

With 302 million active monthly users, Twitter is one of the most popular social networks in the world.¹ In the U.S. alone there are 69.46 million users. The huge amount of data generated by U.S. users is an ideal repository for mining U.S. perceptions of China. It enables us to achieve what have so far evaded opinion polls. Specifically, with a large data set, we are able to carry out state level analysis as never before. By utilizing the time-stamps contained in tweets, we are able to control for fixed time effects. Utilizing the location information, we are able to control for fixed state effects.



Figure 1: America Tweets China, aggregated at the provincial level. This map is generated with 25,677 China-focused tweets. Data does not include Hong Kong SAR, Macao SAR or Taiwan.

Empirically, our study successfully replicates the stylized findings from conventional opinion polls as well as generate

¹https://about.twitter.com/company.

new insights. We find that New York, Michigan, Indiana and Arizona are the top four China-friendly U.S. states and that Wyoming, Wisconsin, South Dakota and West Virginia are the least friendly. Wyoming, South Dakota, Kansas and Nevada are the most homogeneous in attitudes towards China. Michigan, New Hampshire, New Jersey and Wisconsin are the least homogeneous. At the individual level, we find attitudes towards China improve as the individuals' Twitter experience grows longer and more intense. We also find that individuals of Chinese ethnicity are statistically more China-friendly.

Our paper proceeds as follows. Section 2 presents related literature on international relations, sentiment analysis using Twitter data, and inferring geo-information in the tweets. Section 3 presents our data and data processing procedures. Section 4 presents our state-level analysis. Section 5 presents the individual level analysis. Section 6 concludes.

II. LITERATURE REVIEW

Our study builds on previous research both in international relations and in computer science.

China's rise is quickly reshaping the post-Cold War international structure [6]. Lake argues that if China continues to grow, it is likely to bid for its own subordinates to counter America's current hierarchies [7]. Mearsheimer, a proponent of offensive realism, contends that China's rise will not be peaceful and that China will "try to dominate Asia the way the United States dominates the Western Hemisphere" [8].

Perceptions matter in international relations [1], [2] and particularly so in the relations between the U.S. and China. Johnson, through analyzing Chinese publications, argues that the common description in U.S. media, pundit, and academia of an increasingly assertive China is ill-founded and points out the dangers of misperception [9]. Accurate perceptions are likely to contribute to trust while misguided perceptions could well lead to conflict and even war.

We believe tweets can be used to measure U.S. perceptions. The abundance of data generated through Twitter has attracted researchers from various fields. Bollen et al. use Twitter mood to predict the stock market [10]; Paul and Drezde mine tweets for public health topics [11]; and An et al. use Twitter data to track opinions about climate change [12]. Among this group of researchers there are also political scientists. For example, Tumasjan et al. use the tweets to predict election results [13]. Barberá analyzes the network structure of Twitter users to infer political ideology [14].

In order to perform state level analysis, our study makes extensive use of the geo-information in the tweets. In this effort, we benefit from the research by Hecht et al. [15]. Their work identifies various problems with geo-information. We adhere to their suggestions and select only those tweets that have a case-sensitive state address such as *CA* in "Los Angeles, CA", *New York* in "Upper Manhattan, New York" and *Vermont* in "Vermont, USA." For the purpose of this study, we stop at the state level and do not go to the city level.

III. DATA AND PROCESSING

In this section, we first describe our data set and the processing procedures. Second, we define tweet level, state level and user level features respectively. Third, we use timeseries tweets and pseudo-labeled tweets to test the validity of the sentiment analysis tool TextBlob.

A. Data

We compile a corpus of tweets using the Twitter search API between 10th and 29th of March, and between May 16th and June 15th. We perform a query-based search (*China OR Chinese*) to collect English tweets related to China. We then select those tweets with a state address. When a tweet can be attributed to multiple U.S. states, we attribute that tweet to all the identified states. Details of the processing procedure are described below. In the end, after removing duplicates, we have collected 724,146 tweets, each associated with a unique tweet id, a state id, and a time stamp.²

...

Input: tweet_i

Output : tweet _i 's attributes, or null
If tweet _i does not contain stop words
For each state _j \in {51 U.S. states} ³ .
If state _j in tweet _i .place
assign tweet _i to state _j
extract all the attributes
evaluate polarity of tweet _i
For each province _k \in {31 Chinese provinces}
If province _k in tweet _i .text
assign tweet _i to province _k
For each name ₁ \in {100 common Chinese names}
If name ₁ in tweet _i .username
Chinese=1
Else
Chinese=0
End if
End for
End if

From China's Xinhua News Agency's website, we obtain the top 100 most common Chinese surnames, which are used by 84.77% of the Chinese population.⁴ We use these names to identify Twitter users of Chinese ethnicity. When translated into *pinyin*, these 100 names result in 85 distinct names. We report these names in alphabetic order in Table 1.

²The data sets and the codes are available at the authors' website: https://sites.google.com/site/wangyurochester/.

³For the purpose of this study, we treat Washington, D.C. as a state. ⁴http://news.xinhuanet.com/society/2007-04/24/content_6021482.htm.

Table I: Most Common Chinese Surnames

Bai	Cai	Cao	Ceng	Chen				
Cheng	Cui	Dai	Deng	Ding				
Dong	Du	Duan	Fan	Fang				
Feng	Fu	Gao	Gong	Gu				
Guo	Han	Hao	He	Hou				
Hu	Huang	Jia	Jiang	Jin				
Kong	Lei	Li	Liang	Liao				
Lin	Liu	Long	Lu	Luo				
Lv	Ma	Mao	Meng	Mo				
Pan	Peng	Qian	Qin	Qiu				
Ren	Shao	Shi	Song	Su				
Sun	Tan	Tang	Tao	Tian				
Wan	Wang	Wei	Wu	Xia				
Xiang	Xiao	Xie	Xiong	Xu				
Xue	Yan	Yang	Yao	Ye				
Yin	Yu	Yuan	Zhang	Zhao				
Zheng	Zhong	Zhou	Zhu	Zou				

Out of the 101,907 individuals in the sample data set, we are able to identify 938 individuals with Chinese surnames, which represents 0.923% of the sample. This is close to the official figure 1.02%, published by the Census Bureau in 2010.⁵

B. Features

Tweet features:

Polarity: This is defined as how positive the tweet is, ranging from -1 to 1. Polarity is calculated using Textblob. With polarity, we can view each China-focused tweet as a vote.

Other tweet features include followers, followees, retweets, and reply (binary).

State level features:

Friendliness: This is defined as the arithmetic average of the tweets' polarity scores for each state, ranging from -1 to 1. $F_s = \frac{\sum f_{s,i}}{n}$. It measures the state's aggregate sentiment towards China.

Variance: This is defined as the variance of the tweets' polarity scores for each state. $V_s = Var(f_{s,i})$. Variance measures how varied each state's attitudes are towards China. We call a state homogeneous if the variance is small.

Individual level features:

Experience: This is defined as the length of the period the individual has been using Twitter. It is calculated as $date_1$ - $date_0$, where $date_1$ is the day when the tweet is posted and $date_0$ is the day when the Twitter account is created.

Intensity: This is defined as the average number of tweets the individual posts per day. It is calculated as $\frac{\#tweets}{date_1 - date_0 + 1}$.

Chinese: This is defined as whether the individual has a Chinese surname. It is binary.

C. External Validity

We validate the viability of using Textblob to measure sentiments towards China.⁶ So far as we know, there have been no state-level (cross-sectional) opinion polls on U.S.-China relations. All the data measuring U.S. attitudes towards China are limited to the national level. Indeed, this is one of the motivations for this study. We decide to use time series data for the purpose of validation, as significant events between the two countries are easily recognizable and unanimity is easy to achieve.

Three events stood out between May and June: the South China Sea crisis starting on May 20th, the Yangtze River accident on June 1st and Hong Kong protests on June 14th. As shown in Fig. 2 below, these events are well reflected in the aggregate national sentiments. Thus, Textblob passes our time-series test.



Figure 2: U.S. sentiment towards China, May 14-June 15.

We further test Textblob's performance with tweets that contain emoticons. Davidov et al. have shown that smileys, as well as hashtags, in tweets can be used as labels [16]. We first choose two specific smileys: ":)" for positve and ":(" for negative. We are able to find 1255 tweets with these emoticons in our sample. The testing rule is as follows: Correct:

if Textblob returns polarity>=0 for ":)"

or if Textblob returns polarity<=0 for ":("

This is a lenient test as tweets that contain emoticons and are marked as neutral are automatically classified as correct. The testing results, reported below, are satisfactory.

1255
92.4
95.6
72.3

⁶http://textblob.readthedocs.org/en/dev.

⁵http://www.census.gov/prod/cen2010/briefs/c2010br-11.pdf.

IV. STATE LEVEL ANALYSIS

In this section, we investigate the state characteristics of the tweets. We first calculate the volume of tweets that can be attributed to each state and compare our statistics with Google Trend. We then create a **State-Province Matrix** that projects tweets generated in a U.S. state to a Chinese province. The dimension of our matrix is 51 (states) x 31 (provinces). Third, we evaluate the friendliness and variance of each state based on our data set.

A. Volume of Tweets by State

Following the processing procedure described in Section 3, we obtain 724,146 China-focused tweets geocoded to the state level. We then calculate the total number of tweets assigned to each state. The summary statistics are reported below.

Table II: Summary statistics per state

Variable	Mean	Std. Dev.	Min.	Max.	Ν
tweets	14199	23085	1836	149043	51

In terms of the total number of tweets, the top four states are New York, California, Washington, D.C., and Texas. When controlling for state population, Washington, D.C. generates by far the most China-focused tweets per capita. The bottom four states are Vermont, New Mexico, West Virginia and South Dakota. Detailed geographical comparisons are reported in Fig. 3.



Figure 3: America tweets China, aggregated at state level. This map is generated with 724,146 China-focused tweets. Washington, D.C. is not shown on the map.

For cross validation, we compare our results with the state-based index generated from the Google Trend.⁷ The Google index measures the frequency with which people in each U.S. state search for the keyword *China* between Jan. 4, 2004 and Jun. 21, 2015. To make for easy comparison, we

first log-transform our counts of tweets. The Google index is used as it is. We plot the Google index as the x axis and plot our Twitter index as the y axis. The result is reported in Fig. 4.

The Twitter index and the Google index are highly correlated, with a correlation coefficient of 0.76 (sample size: 51). One pattern stands out here. The top three states New York, California and Washington, DC score very high by both measures. The bottom four states South Dakota, Montana, Wyoming and New Mexico score very low by both measures. The remaining forty-four states lie in between.



Figure 4: Compare Twitter with Google. β represents the linear regression coefficient. Standard error of the estimation is reported in parentheses.

B. The State-Province Matrix

The large size of our data set enables us to achieve state level analysis that has so far evaded most researchers. Moreover, by examining the contents of these tweets more closely, we are able to identify which Chinese province a tweet is targeted at. Connecting the state of origin to the province of destination, we can build a State-Province matrix of dimension 51×31 . The matrix, built with 25,677 tweets, is reported in Table 3. Each row represents a U.S. state and each column a Chinese province. Values in each row represent the distribution of tweets in the 31 Chinese provinces. Their sum has been normalized to 1.

Two immediate observations follow. First, nationwide most of the tweets can be attributed to three Chinese provinces: Beijing, Shanghai and Tibet. Xinjiang province makes a distant fourth. Second, there exists large inter-state variation. For example, the share of tweets that go to Beijing ranges from 24.3% for Delaware to 60.0% for Rhode Island. For the majority of the U.S. states (44 out of 51), the largest share of tweets goes to Beijing.⁸ For Delaware, Idaho, New

⁷https://www.google.com/trends/

⁸We observe two ties and in both cases we decide to side with Beijing.

Table III: State-Province Matrix

Matrix	皖	京	渝	闽	甘	粤	桂	贵	琼	冀	黑	豫	鄂	湘	蒙	苏	赣	吉	辽	宁	青	陕	鲁	沪	晋	JI]	津	藏	新	굸	浙
AK	1.5	29.2	1.5	0.0	0.0	0.0	0.0	0.0	0.0	1.5	0.0	3.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	23.1	0.0	4.6	0.0	29.2	4.6	1.5	0.0
AL	0.3	39.5	1.3	0.0	0.0	1.0	3.5	0.3	0.8	1.0	0.0	1.8	1.8	1.3	0.0	0.5	0.3	0.0	0.0	0.0	0.3	0.0	0.3	18.9	0.0	2.5	3.8	16.6	2.5	1.8	0.3
AR	0.3	41.1	1.7	0.0	0.9	2.0	2.3	0.0	0.3	0.6	0.0	0.6	1.4	1.1	0.6	0.3	0.0	0.3	0.6	0.0	0.3	0.9	0.0	22.6	0.0	1.1	2.3	14.0	3.4	1.4	0.0
AZ	0.9	50.0	0.9	0.9	0.0	2.7	2.7	0.6	1.8	0.0	0.0	1.2	0.9	0.6	0.3	0.0	0.3	0.0	0.6	0.0	1.2	0.0	0.3	19.3	0.0	1.5	1.8	8.4	1.8	1.2	0.0
CA	0.3	37.0	1.0	0.2	0.2	0.8	14.6	0.4	1.6	0.6	0.1	0.9	1.0	2.0	0.0	0.5	0.1	0.1	0.4	0.2	0.2	0.2	0.5	19.9	0.2	1.7	0.9	10.3	2.7	1.3	0.3
CO	0.4	35.3	1.7	0.0	0.2	0.4	2.1	0.4	0.6	0.6	0.0	1.3	1.3	1.9	0.0	0.4	0.2	0.2	0.6	0.0	0.4	0.0	0.2	19.3	0.0	1.5	0.6	14.1	13.1	2.8	0.0
CT	1.1	36.7	0.7	0.4	0.0	2.6	0.7	0.0	1.5	4.9	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.9	25.1	0.0	1.5	4.1	13.9	3.7	0.4	0.4
DC	0.1	50.9	1.0	0.5	0.1	0.8	0.4	0.1	1.5	0.5	0.1	0.5	0.7	0.5	0.1	1.0	0.3	0.2	1.0	0.1	0.5	0.4	0.5	15.2	0.2	0.7	1.0	12.0	7.6	1.2	0.5
DE	0.7	24.3	0.7	0.7	1.4	0.7	0.7	0.7	0.7	1.4	0.0	1.4	1.4	0.7	0.0	0.7	0.0	0.0	0.7	0.7	0.7	0.7	0.7	26.4	0.0	2.0	0.0	21.6	9.5	0.7	0.7
FL	0.1	43.3	1.1	0.3	0.1	0.3	2.5	0.0	1.8	0.3	0.2	1.3	1.0	1.4	0.1	0.1	0.1	0.0	0.2	0.1	0.2	0.1	0.3	30.7	0.2	1.7	1.1	7.6	1.1	2.2	0.6
GA	0.0	52.6	0.2	2.0	0.2	0.2	2.2	0.4	0.9	0.7	0.4	1.1	1.5	1.5	0.4	0.7	0.0	0.2	0.7	0.0	0.0	0.0	0.2	17.4	0.2	0.2	0.2	12.0	2.4	1.1	0.4
HI	0.3	58.9	0.6	0.0	0.3	0.6	0.9	0.3	1.2	0.0	0.0	0.9	0.0	0.0	0.3	0.3	0.0	0.0	0.3	0.0	0.9	0.9	0.0	14.8	0.0	0.9	0.3	11.8	3.0	1.2	1.2
IA	0.0	41.9	1.3	0.0	0.0	0.0	4.4	0.0	1.3	1.3	0.0	0.6	1.3	1.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	23.1	0.0	2.5	0.0	12.5	6.3	1.9	0.6
ID	0.0	29.6	0.7	0.0	0.7	0.0	0.7	0.7	0.0	0.0	0.0	2.1	0.7	1.4	0.0	0.7	0.0	0.0	0.0	0.0	0.7	0.0	0.0	38.0	0.7	1.4	0.7	16.2	2.1	2.8	0.0
IL	0.5	46.9	0.5	0.5	0.2	0.5	2.3	0.7	0.7	1.6	0.0	0.9	0.5	0.9	0.0	0.2	0.2	0.5	0.2	0.0	0.2	0.0	0.5	21.4	0.0	2.3	1.9	12.4	1.4	0.9	0.9
IN	0.6	39.4	0.6	0.5	0.0	1.0	2.1	0.6	1.7	0.8	0.1	1.4	0.4	1.2	0.3	0.6	0.2	0.5	0.1	0.0	0.1	0.1	0.6	22.5	0.1	1.6	1.0	15.1	4.0	2.5	0.6
KS	0.0	41.9	0.5	1.0	0.0	1.4	1.4	0.0	1.0	0.0	0.0	0.0	1.4	1.4	0.0	0.0	0.0	0.5	1.0	0.0	0.0	0.0	0.0	32.9	0.0	0.0	0.5	7.1	6.7	1.4	0.0
KY	0.0	31.3	2.4	0.0	0.0	1.2	1.8	1.2	0.6	0.0	0.0	1.2	1.2	0.0	0.0	1.2	0.0	0.6	6.0	0.0	0.0	0.0	0.0	16.3	0.0	0.0	1.2	25.3	4.2	4.2	0.0
LA	0.0	29.2	0.9	0.9	0.3	0.3	0.9	0.9	2.4	0.6	0.0	1.2	0.6	2.1	0.3	0.0	0.6	0.0	0.0	0.0	0.3	0.0	0.0	21.1	0.0	2.1	0.9	32.2	0.9	1.2	0.0
MA	0.0	34.7	0.7	1.2	0.0	0.5	3.0	0.7	0.9	0.5	0.2	9.3	1.2	0.5	0.0	0.0	0.0	0.0	1.2	0.0	0.0	0.0	0.2	27.4	0.0	2.1	2.3	9.1	1.6	2.1	0.7
MD	0.4	44.1	1.3	0.0	0.0	0.0	2.6	0.0	1.8	1.8	0.0	0.4	0.9	0.4	0.0	0.0	0.0	0.0	0.4	0.0	0.9	0.4	0.0	19.8	0.0	3.1	0.9	13.7	5.7	0.9	0.4
ME	0.0	36.6	0.0	0.0	0.0	1.0	1.0	0.0	2.1	0.5	0.0	0.0	1.0	2.1	0.0	0.5	0.5	0.0	1.0	0.0	0.0	0.0	1.0	30.9	0.0	1.6	2.6	11.0	4.2	2.1	0.0
MI	0.2	54.7	0.5	0.3	0.0	0.3	1.6	0.0	0.5	0.2	0.2	1.0	0.6	1.1	0.0	0.3	0.0	0.3	0.2	0.0	0.0	0.0	0.3	21.2	0.0	1.5	2.3	10.7	0.6	1.1	0.3
MN	1.0	38.3	0.0	0.0	0.5	1.0	2.5	0.5	2.0	0.5	0.0	0.5	0.5	0.5	0.0	0.5	0.0	0.5	0.5	0.0	0.0	0.0	0.0	25.4	0.0	2.0	4.5	14.9	2.5	1.5	0.0
MO	0.0	40.0	0.0	1.6	0.0	1.6	2.3	0.0	1.6	0.8	0.0	1.6	2.3	0.8	1.6	0.8	0.0	0.0	0.8	0.0	0.8	0.0	0.0	22.7	0.0	0.8	2.3	9.4	3.1	4.7	0.0
MS	0.0	37.7	0.8	0.0	0.0	0.0	0.8	0.0	0.0	1.6	0.0	2.5	3.3	3.3	0.0	0.0	0.0	0.0	2.5	0.0	0.0	0.0	0.0	24.6	0.0	0.8	0.0	15.6	0.0	5.7	0.8
MI	0.0	31.5	1.5	1.3	0.0	0.0	0.3	0.0	2.5	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.8	0.0	0.0	0.0	1.5	25.0	0.0	1.5	0.0	13.8	1.5	1.5	0.0
NC	0.2	30.5	0.9	0.0	0.2	1.1	2.1	0.2	9.1	0.2	0.0	0.0	1.5	1.1	0.0	0.2	0.0	0.0	0.0	0.0	0.2	0.0	0.4	28.7	0.0	0.4	0.2	14.2	0.9	0.4	0.2
NE	0.0	30.3 25 7	0.4	0.0	0.0	0.8	3.4 2.1	1.1	0.8	0.8	0.4	1.5	0.4	0.8	0.0	0.8	0.0	0.0	0.0	0.0	0.4	0.0	0.4	23.3	0.0	1.1	1.5	21.4 11.6	1.9	2.3	0.4
NL	0.0	35.7	0.9	1.0	0.0	1.5	5.1	0.9	1.0	0.4	0.0	24.4	0.9	0.9	0.4	0.9	0.0	0.0	0.9	0.0	0.4	0.0	0.4	27.2	0.0	1.0	0.9	7.0	2.2	4.9	0.4
NI	1.6	51.5	1.2	0.2	0.0	0.0	1.7	0.0	1.0	0.8	0.0	2 4.4	0.8	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.5 17.9	0.0	1.0	0.0	9.2	2.4	1.6	0.0
INJ	1.0	54.9 10.2	1.2	0.5	0.2	0.5	1./	0.9	2.4	0.7	0.0	0.5	0.7	0.9	0.0	0.5	0.0	0.0	0.9	0.0	0.2	0.2	0.7	17.0	0.0	1.2	6.8	0.3	1./	1.0	0.5
NV	1.1	46 1	0.6	0.0	1.1	0.0	2.4	1.1	1.1	1.1	0.0	0.6	1.7	2.3	0.0	0.6	0.0	0.0	1.1	0.0	0.0	0.0	0.0	16 1	0.0	0.6	0.6	13.0	3.4	0.6	2.2
NV	0.7	40.1	0.0	0.0	1.1	0.0	0.6	0.3	0.7	0.3	0.0	0.0	0.5	1.2	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	21.8	0.0	1.0	0.0	16.2	17	0.0	0.2
ОН	0.7	37 3	0.7	0.7	0.0	0.9	2.2	0.7	13	0.5	0.1	17	0.5	1.2	0.0	0.2	0.0	0.0	0.2	0.0	0.1	0.2	0.2	21.0	0.0	4.1	1.5	7.6	6.8	3.1	1.1
OK	0.0	35.4	0.7	0.0	0.0	0.5	1.2	0.0	0.6	0.4	0.0	1.7	1.9	0.6	0.0	0.2	0.0	0.0	0.5	0.0	0.6	0.0	0.0	29.8	0.0	1.1	0.6	15.5	43	1.9	1.1
OR	0.3	32.5	2.4	0.8	0.3	1.8	2.1	0.3	0.8	1.3	0.0	0.8	1.0	1.0	0.3	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.5	32.5	0.3	2.4	1.3	12.6	3.1	1.6	0.0
PA	0.5	29.3	1.0	0.3	0.2	0.7	2.0	0.7	0.7	0.2	0.2	1.0	0.5	0.3	0.0	0.0	0.0	0.3	0.5	0.0	0.0	0.2	0.3	20.0	0.0	1.5	3.0	16.9	17.9	1.2	1.0
RI	0.4	60.0	0.0	0.0	0.0	0.7	1.1	0.4	0.7	0.4	0.0	1.1	1.4	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.4	20.4	0.4	0.0	4.2	3.9	2.1	2.1	0.4
SC	0.5	49.5	1.1	0.0	0.0	1.6	1.6	0.5	2.1	0.0	0.0	1.6	1.1	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	23.9	0.5	0.5	0.5	10.6	2.1	1.6	0.0
SD	0.0	31.3	2.1	0.0	0.0	2.1	0.0	0.0	0.0	2.1	0.0	0.0	0.0	2.1	0.0	0.0	0.0	0.0	2.1	0.0	0.0	0.0	0.0	43.8	0.0	2.1	2.1	6.3	4.2	0.0	0.0
TN	0.0	52.4	1.0	0.0	0.0	1.0	1.0	0.5	0.5	0.5	0.0	0.5	1.0	1.4	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	19.2	0.0	2.4	1.9	10.6	2.4	1.4	0.5
TX	0.4	42.0	0.8	0.0	0.1	1.0	3.0	0.3	0.4	1.5	0.4	8.1	1.1	1.9	0.1	0.5	0.2	0.0	0.4	0.0	0.3	0.0	0.4	22.4	0.1	1.9	1.1	8.9	1.3	1.3	0.4
UT	1.1	39.5	2.6	0.0	0.0	0.0	2.1	1.1	1.1	2.1	0.0	0.5	2.1	1.6	0.5	0.5	1.1	0.5	1.6	0.0	0.5	0.5	1.1	23.2	0.5	2.6	1.1	7.4	1.1	2.1	2.1
VA	0.5	49.0	1.6	0.5	0.3	0.3	2.7	0.3	0.5	0.5	0.0	0.5	1.1	1.9	0.0	1.1	0.0	0.3	0.5	0.0	0.0	0.0	0.0	22.5	0.3	0.5	1.1	9.3	1.9	2.5	0.0
VT	0.0	28.6	0.0	0.0	0.0	0.0	1.3	0.0	0.0	0.0	0.0	0.0	1.3	0.0	0.0	0.0	0.0	0.0	1.3	0.0	1.3	0.0	0.0	32.5	0.0	1.3	0.0	15.6	11.7	5.2	0.0
WA	0.3	42.1	1.1	0.9	0.2	0.5	0.9	0.5	2.6	0.9	0.0	0.6	0.6	0.9	0.3	0.8	0.2	0.3	0.5	0.0	0.2	0.0	0.8	24.8	0.2	1.8	0.8	13.4	1.8	2.3	0.2
WI	1.0	41.1	1.0	2.4	0.3	1.0	2.1	0.3	1.0	0.0	0.0	2.4	1.4	2.1	0.7	0.7	0.3	0.0	0.0	0.0	0.0	0.0	0.3	22.6	1.4	2.1	1.4	8.6	3.4	2.1	0.3
WV	0.0	29.9	0.0	1.5	0.0	1.5	3.0	0.0	0.0	1.5	0.0	0.0	1.5	0.0	1.5	1.5	0.0	1.5	3.0	0.0	0.0	0.0	0.0	25.4	0.0	4.5	1.5	16.4	4.5	1.5	0.0
WY	0.0	32.6	2.2	0.0	0.0	0.0	4.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.2	2.2	0.0	0.0	4.3	4.3	0.0	34.8	0.0	0.0	2.2	6.5	4.3	0.0	0.0

Note: (in original order)

皖: Anhui, 京:Beijing, 渝: Chongqing, 闽: Fujian, 甘: Gansu, 粤: Guangdong, 桂: Guangxi, 贵: Guizhou, 琼: Hainan,

冀: Hebei, 黑: Heilongjiang, 豫: Henan, 鄂: Hubei, 湘: Hunan, 蒙: Inner Mongolia, 苏: Jiangsu, 贛: Jiangxi, 吉: Jilin,

辽: Liaoning, 宁: Ningxia, 青: Qinghai, 陕: Shaanxi, 鲁: Shandong, 沪: Shanghai, 晋: Shanxi, 川: Sichuan, 津: Tianjin,

藏: Tibet, 新: Xinjiang, 云: Yunnan, 浙: Zhejiang.

Mexico, South Dakota, Vermont and Wyoming (6 out of 51), Shanghai receives the most tweets. For Louisiana, it is Tibet.

C. Friendliness and Variance

Opinion polls on U.S.-China relations mostly stop at the national level. By contrast, we are able to explore state level nuances. In particular, we are able to measure the friendliness of each state towards China. This is important because hardly ever have international relations been measured without regard to security issues, and security concerns vary across countries. Studying state level attitudes enables us to control for security issues in a perfect way as we can assume all U.S. states share the same security concerns with regards to China. In addition to friendliness, we are also able to measure how varied perceptions of China are in each state.

The results are reported in Fig. 5. We find that the top four most friendly states are New York, Michigan, Indiana and Arizona and that the least friendly four states are Wyoming, Wisconsin, South Dakota and West Virginia. In terms of variance, we find the top four most homogeneous states are Wyoming, South Dakota, Kansas and Nevada and that the least homogeneous states are Michigan, New Hampshire, New Jersey and Wisconsin.

Though we do not explore it here, we point out that it will be of great significance to analyze the causal mechanisms behind these differences between states. We suggest that international trade and Chinese immigration can be influential factors.

V. INDIVIDUAL LEVEL ANALYSIS

In this section, we first introduce our statistical model and then report our main estimation results. Lastly, we perform two separate F tests on the state and date control variables.

A. The Model

Our primary goal here is to replicate some of the stylized results from opinion polls [17], [3]:

- American individuals who follow news about China have less favorable views of China.
- American opinion leaders are more likely to have a favorable view of China.
- American students who have studied in China are more likely to have a favorable view of China.

Our dependent variable is the polarity of the tweet. We use **followers** and **intensity**, introduced in Section 2, to capture the effects of opinion leaders. That is, we treat individuals with more followers and more posted tweets as opinion leaders. In the original study, the opinion leaders consist of U.S. government officials, think tank leaders, media personnel, business executives, and university faculties [17]. We use variables **followees** and **retweet** to capture the effects of following news. The variable **reply** controls for the effects of being in a conversation.

Additionally, we examine the effects of Chinese ethnicity on attitudes towards China. The size of our data set also allows us to control for state and time fixed effects. As previous sections have shown, there is large variation across U.S. states and between different days, so we decide to control for both state and time effects. Altogether, this suggests that following statistical model:

$$y_{i} = \beta_{1} \cdot followers_{i} + \beta_{2} \cdot followees_{i} + \beta_{3} \cdot retweet_{i} + \beta_{4} \cdot reply_{i} + \beta_{5} \cdot experience_{i} + \beta_{6} \cdot intensity_{i} + D_{i}\boldsymbol{\alpha} + C_{i}\boldsymbol{\lambda} + S_{i}\boldsymbol{\gamma} + \epsilon_{i}$$

where $D_i = [d_1^i d_2^i \dots d_{17}^i]$ control for date effects, $S_i = [s_1^i s_2^i \dots s_{50}^i]$ control for state effects and C_i controls for the effects of Chinese ethnicity.

B. The Main Results

We estimate our model with OLS regression and the estimation results are reported in Table 4. Each column represents one regression and has its own specification. The first column displays estimates without controlling for state and time fixed effects. The second column presents the same coefficients but controls for state fixed effects. The third column controls for time fixed effects. The fourth column incorporates both state and time fixed effects.

For the fifth column, we aggregate the tweets on an individual and daily basis by taking the average. This reduces the number of observations from 245,664 to 162,982. The six column, with 242,673 observations, reports estimates for individuals not identified as ethnically Chinese. The seventh column, with 2991 observations, reports estimates for individuals identified as ethnically Chinese.

Our results (Columns 1-5) are consistent with the findings cited above. Specifically, the coefficient for **intensity** is positive and statistically significant, indicating that opinion leaders, in our case individuals who tweet more often, are more likely to have a positive view of China. The coefficient for **followers** is not statistically significant.

Coefficients for **followees** and **retweet** are both negative and statistically significant. This is consistent with the finding that individuals who follow news about China have less favorable views of the country.

We also find **experience** to have a statistically positive effect. This suggests a positive learning experience using Twitter and supports that finding that American students who have studied in China tend to have more positive views of the country.

The estimate for **Chinese** is positive in all the first five specifications, but is statistically significant only in Test 5, when we aggregate individuals' daily tweets.

Comparing the results in the sixth column and the seventh column, we find individuals of Chinese ethnicity behave differently from the rest of the sample. For example, the coefficient on **followers** is positive and statistically significant.



Figure 5: Friendliness and Variance. The top figure reports the friendliness index for each state. The bottom figure reports the variance index for each state.

				tweet polarity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Baseline	State effects	Date effects	State & Date effects	Aggregated	Non-Chinese Ethnicity	Chinese Ethnicity
followers	2.80e-11	-1.33e-09	1.62e-10	-1.21e-09	-9.00e-11	-1.23e-09	0.000000725**
	(0.02)	(-0.79)	(0.10)	(-0.72)	(-0.04)	(-0.73)	(2.73)
followees	-0.00000353***	-0.000000333***	-0.000000352***	-0.000000331***	-0.000000154**	-0.000000326***	-0.00000138**
	(-8.22)	(-7.74)	(-8.19)	(-7.70)	(-2.80)	(-7.52)	(-3.04)
retweet	-0.000000907**	-0.000000871**	-0.000000935**	-0.000000897**	-0.00000154***	-0.000000898**	0.00000340
	(-2.94)	(-2.83)	(-3.03)	(-2.91)	(-4.75)	(-2.91)	(0.18)
reply	0.0219***	0.0217***	0.0219***	0.0217***	0.0219***	0.0215***	0.0192
	(17.10)	(16.90)	(17.11)	(16.91)	(14.41)	(16.64)	(1.73)
experience	0.00000337***	0.00000293***	0.00000332***	0.00000287***	0.00000170*	0.00000290***	-0.000000992
	(5.68)	(4.91)	(5.60)	(4.81)	(2.46)	(4.82)	(-0.17)
intensity	0.0000930***	0.0000870***	0.0000927***	0.0000865***	0.0000855***	0.0000869***	-0.000113
	(78.96)	(67.11)	(78.46)	(66.54)	(40.29)	(66.62)	(-1.85)
Chinese	0.000187	0.00337	0.000348	0.00344	0.0210***		
	(0.05)	(0.83)	(0.09)	(0.85)	(3.98)		
state effects	No	Yes	No	Yes	Yes	Yes	Yes
date effects	No	No	Yes	Yes	Yes	Yes	Yes
constant	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	245664	245664	245664	245664	162982	242673	2991
adj. R^2	0.028	0.030	0.030	0.031	0.016	0.031	0.042

Table IV: Estimation Results

t statistics in parentheses * p < 0.05, ** p < 0.01, *** p < 0.001

Estimates for control variables are not reported. In the following subsection, joint F tests are used to show that they are statistically different from zero.

The estimates on **retweet**, **reply**, **experience** and **intensity** are no longer statistically significant.

C. F tests

To test the significance of the state and date effects, we perform two separate F tests based on the fourth column in Table 4. The testing results are reported below. With F(50, 245589)=6.88, we reject the null hypothesis that coefficients on state control variables are all zeros. With F(17, 245589)=16.99, we are able to reject the null hypothesis that the coefficients on date control variables are all zeros. These tests further confirm the existence of inter-state and inter-day variations.

Table V: F Tests on control variables

state coefficients	F(50, 245589)=6.88	Prob>F=0.0000
date coefficients	F(17, 245589)=16.99	Prob>F=0.0000

VI. CONCLUSION

The U.S.-China relationship is arguably the most important bilateral relationship in the 21st century. It demands detailed measurement and analysis. In this paper, we have proposed a new method to measure this relationship using tweets. With a large data set, we are able to carry out state level analysis. Utilizing geo-information and time stamps, we can control for fixed state and time effects. We demonstrate the existence of inter-state and inter-day variations and control for them in our regression analysis.

Our work replicates some stylized results from opinion polls as well as generate new insights. At the state level, we find New York, Michigan, Indiana and Arizona are the top four most China-friendly states. Wyoming, South Dakota, Kansas and Nevada are most homogeneous. At the individual level, we find attitudes towards China improve as the individuals' Twitter experience grows longer and more intense. We also find individuals of Chinese ethnicity are more China-friendly.

Our study is the first to analyze U.S.-China relations using Twitter data. The results we achieve are very encouraging. In future research we intend to increase the size of our data set and further refine our analytic tools.

ACKNOWLEDGMENT

Yu Wang would like to thank the Department of Political Science at the University of Rochester for warm encouragement and generous funding. This work was also generously supported in part by Google, Yahoo, Adobe, TCL, and New York State CoE CEIS and IDS.

REFERENCES

[1] R. Jervis, *Perceptions and Misperceptions in International Politics.* Prince, 1976.

- [2] A. H. Kydd, *Trust and Mistrust in International Relations*. Princ, 2005.
- [3] (2015, students china, May) For american in some risks, no regrets. Foreign Policy. Available: http://foreignpolicy.com/2015/05/27/ [Online]. american-students-in-china-fp-survey-no-regrets-china-u-cross-cultural-nati
- [4] A. Dugan. (2014, February) Americans view china mostly unfavorably. [Online]. Available: http://www.gallup.com/ poll/167498/americans-view-china-mostly-unfavorably.aspx? version=print
- [5] J. R. Lax and J. H. Phillips, "How should we estimate public opinion in the states," *American Joural of Political Science*, vol. 53, no. 1, pp. 107–121, January 2009.
- [6] K. N. Waltz, *Theory of International Politics*. McGraw-Hill Publishing Company, 1979.
- [7] D. A. Lake, *Hierarchy in International Relations*. Cornell University Press, 2009.
- [8] J. J. Mearsheimer, *The Tragedy of Great Power Politics*. W.W.Norton & Company, Inc., 2014.
- [9] A. I. Johnson, "How new and assertive is china's new assertiveness?" *International Security*, vol. 37, no. 4, pp. 7– 48, 2013.
- [10] J. Bollen, H. Mao, and X.-J. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, March 2011.
- [11] M. J. Paul and M. Dredze, "You are what you tweet: Analyzing twitter for public health." in *ICWSM*, L. A. Adamic, R. A. Baeza-Yates, and S. Counts, Eds. The AAAI Press, 2011.
- [12] X. An, A. R. Ganguly, Y. Fang, S. B. Scyphers, A. M. Hunter, and J. G. Dy, "Tracking climate change opinions from twitter data." Workshop on Data Science for Social Good, 2014.
- [13] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," in *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 2010, pp. 178–185.
- [14] P. Barbera, "Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data," *Political Analysis*, vol. 23, pp. 76–91, September 2014.
- [15] B. Hecht, L. Hong, B. Suh, and E. H. Chi, "Tweets from justin bieber's heart: the dynamics of the location field in user profiles." in *CHI*, D. S. Tan, S. Amershi, B. Begole, W. A. Kellogg, and M. Tungare, Eds. ACM, 2011, pp. 237–246.
- [16] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys," *Coling 2010: Poster Volume*, pp. 241–249, 2010.
- [17] C. English. (2012, February) Americans, opinion leaders see u.s.-china ties as friendly. Gallup. [Online]. Available: http://www.gallup.com/poll/152618/ americans-opinion-leaders-china-ties-friendly.aspx?version= print