arXiv:1302.3299v2 [physics.soc-ph] 19 Feb 2013

The Geography of Happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place

Lewis Mitchell,[1,*] Morgan R. Frank,[1,†] Kameron Decker Harris,[1,‡]
Peter Sheridan Dodds,[1,§] and Christopher M. Danforth[1,¶]

[1] *Department of Mathematics & Statistics, Vermont Complex Systems Center, Computational Story Lab, & the Vermont Advanced Computing Core, The University of Vermont, Burlington, VT 05401.*

(Dated: February 20, 2013)

We conduct a detailed investigation of correlations between real-time expressions of individuals made across the United States and a wide range of emotional, geographic, demographic, and health characteristics. We do so by combining (1) a massive, geo-tagged data set comprising over 80 million words generated over the course of several recent years on the social network service Twitter and (2) annually-surveyed characteristics of all 50 states and close to 400 urban populations. Among many results, we generate taxonomies of states and cities based on their similarities in word use; estimate the happiness levels of states and cities; correlate highly-resolved demographic characteristics with happiness levels; and connect word choice and message length with urban characteristics such as education levels and obesity rates. Our results show how social media may potentially be used to estimate real-time levels and changes in population-level measures such as obesity rates.

I. INTRODUCTION

With vast quantities of real-time, fine-grained data, describing everything from transportation dynamics, resource usage, and social interactions, the science of cities has entered the realm of the data-rich fields. While much work and development lies ahead, the opportunity to scientifically engage with urban phenomena has now become broadly available to quantitatively-minded researchers [5]. And with over half the world's population now living in urban areas, and this proportion continuing to grow, cities, long central to human society, will only become increasingly more so [22]. Our focus here concerns one of the many important questions we are led to continuously address about cities: how does living in urban areas relate to well-being? Such an undertaking is part of a general program seeking to quantify and explain the evolving cultural character—the stories—of cities, as well as geographic places of larger and smaller scales.

Numerous studies on well-being are published every year. The UN's 2012 World Happiness Report attempts to quantify happiness on a global scale using a 'Gross National Happiness' index which uses data on rural-urban residence and other factors [28]. In the US, Gallup and Healthways produce a yearly report on the well-being of different cities, states and congressional districts [19], and maintain a well-being index based on continual polling and survey data [3]. Other countries are beginning to produce metrics measuring well-being: in 2012, surveys measuring national well-being and how it relates to both heath and where people live were conducted in both the United Kingdom by the Office of National Statistics [4, 27] and in Australia by Fairfax Media and Lateral Economics [16].

While these and other approaches to quantifying the sentiment of a city as a whole rely almost exclusively on survey data, there is now a range of complementary, remote-sensing methods available to researchers. The explosion of the amount and availability of data relating to social network use in the past 15 years has driven a rapid increase in the application of data-driven techniques to the social sciences and sentiment analysis of large-scale populations.

Our overall aim in this paper is to investigate how geographic place correlates with and potentially influences societal levels of happiness. In particular, after first examining happiness dynamics at the level of states, we will explore urban areas in the United States in depth, and ask if it is possible to (a) measure the overall average happiness of people located in cities; and (b) explain the variation in happiness across different cities. Our methodology for answering the first question uses word frequency distributions collected from a large corpus of geolocated messages or 'tweets' collected from Twitter, with individual words scored for their happiness independantly by users of Amazon's Mechanical Turk service [2]. This technique was introduced by Dodds and Danforth (2009) [10] and greatly expanded upon in Dodds et al. (2011) [11], as well as tested for robustness and sensitivity. In attempting to answer the second question of happiness variability, we examine how individual word usage correlates with happiness and various social and economic factors. To do this we use the 'word shift graph' technique developed in [10, 11], as well as correlate word usage frequencies with traditional city-level census survey data. As we will show, the combination of these

*Electronic address: lewis.mitchell@uvm.edu; Corresponding author
†Electronic address: morgan.frank@uvm.edu
‡Electronic address: kamerondeckerharris@gmail.com
§Electronic address: peter.dodds@uvm.edu
¶Electronic address: chris.danforth@uvm.edu

techniques produces significant insights into the character of different cities and places.

We structure our paper as follows. In Section II, we describe the data sets and our methodology for measuring happiness. In Section III we measure the happiness of different states and cities and determine the happiest and saddest states and cities in the US, with some analysis of why places vary with respect to this measure. In Section IV we compare our results for cities with census data, correlating happiness and word usage with common economic and social measures. We also use the word frequency distributions to group cities by their similarities in observed word use. We conclude with a discussion in Section V.

II. DATA AND METHODOLOGY

We examine a corpus of over 10 million geotagged tweets gathered from 373 urban areas in the contiguous United States during the calendar year 2011. This corpus is a subset of Twitter's garden hose feed, and represents roughly 10% of all geotagged tweets posted in 2011. Urban areas are defined by the 2010 United States Census Bureau's MAF/TIGER (Master Address File/Topologically Integrated Geographic Encoding and Referencing) database [9]. See Appendix A for a more detailed description of the data set as well as an exploration of the relationship between area and perimeter, or fractal dimension, of these cities.

To measure sentiment (hereafter happiness) in these areas from the corpus of words collected, we use the Language Assessment by Mechanical Turk (LabMT) word list (available online in the supplementary material of [11]), assembled by combining the 5,000 most frequent words occurring in each of four text sources: Google Books (English), music lyrics, the New York Times and Twitter. A total of roughly 10,000 of these individual words have been scored by users of Amazon's Mechanical Turk service on a scale of 1 (sad) to 9 (happy), resulting in a measure of average happiness for each given word [23]. For example, 'rainbow' is one of the happiest words in the list with a score of $h_{\text{avg}} = 8.1$, while 'earthquake' is one of the saddest, with $h_{\text{avg}} = 1.9$. Neutral words like 'the' or 'thereof' tend to score in the middle of the scale, with $h_{\text{avg}} = 4.98$ and 5 respectively.

For a given text T containing N unique words, we calculate the average happiness h_{avg} by

$$h_{\text{avg}}(T) = \frac{\sum_{i=1}^{N} h_{\text{avg}}(w_i) f_i}{\sum_{i=1}^{N} f_i} = \sum_{i=1}^{N} h_{\text{avg}}(w_i) p_i \qquad (1)$$

where f_i is the frequency of the ith word w_i in T for which we have a happiness value $h_{\text{avg}}(w_i)$, and $p_i = f_i / \sum_{i=1}^{N} f_i$ is the normalized frequency of word w_i.

Importantly, with this method we make no attempt to take the context of words or the meaning of a text into account. While this may lead to difficulties in accurately determining the emotional content of small texts, we find that for sufficiently large texts this approach nonetheless gives reliable (if eventually improvable) results. An analogy is that of temperature: while the motion of a small number of particles cannot be expected to accurately characterize the temperature of a room, an average over a sufficiently large collection of such particles defines a durable quantity. Furthermore, by ignoring the context of words we gain both a computational advantage and a degree of impartiality; we do not need to decide *a priori* whether a given word has emotional content, thereby reducing the number of steps in the algorithm and hopefully reducing experimental bias.

Following Dodds et al. (2011), for the remainder of this paper, we remove all words w_i for which the happiness score falls in the range $4 < h_{\text{avg}}(w_i) < 6$ when calculating $h_{\text{avg}}(T)$. Removal of these neutral or 'stop' words has been demonstrated to provide a suitable balance between sensitivity and robustness for our 'hedonometer' [11]. Further details on how we preprocessed the Twitter data set can be found in Appendix A.

We will correlate our happiness results with census data taken from the American Community Survey 1-year estimates for 2011, accessed online at http://factfinder2.census.gov/.

III. HAPPINESS ACROSS STATES AND URBAN AREAS

We first examine how happiness varies on a somewhat coarser scale than we will focus on for the majority of this paper, by plotting the average happiness of all states in the US in figure 1. To avoid the problem that some states have happier names than others (for example, Hawaii), we removed each state name from the calculation for h_{avg}. We remark first that at such a coarse resolution there is little variation between states, which all lie between 0.15 of the mean value for the entire United States of $h_{\text{avg}} = 6.01$. The happiest state is Hawaii with a score of $h_{\text{avg}} = 6.17$ and the saddest state is Louisiana with a score of $h_{\text{avg}} = 5.88$. Hawaii emerges as the happiest state due to an abundance of relatively happy words such as 'beach' and food-related words, but also because of the presence of the word 'hi'. This is most likely because of an increased use of Hawaii's state code 'HI' in geotagged tweets, and will somewhat bias the results. However, we chose not to remove this word from the data set because its use in place of 'hello' will contribute to the happiness score in other states, and the rich variety of happy words occurring in Hawaii paints a convincing picture of it as a happy state regardless of this small bias. A similar result showing greater happiness and a relative abundance of food-related words in tweets made by users who regularly travel large distances (as would be the case for many of the tweets emanating from Hawaii) has been reported in [18]. Louisiana is revealed as the saddest state primarily as a result of an abundance of profanity relative

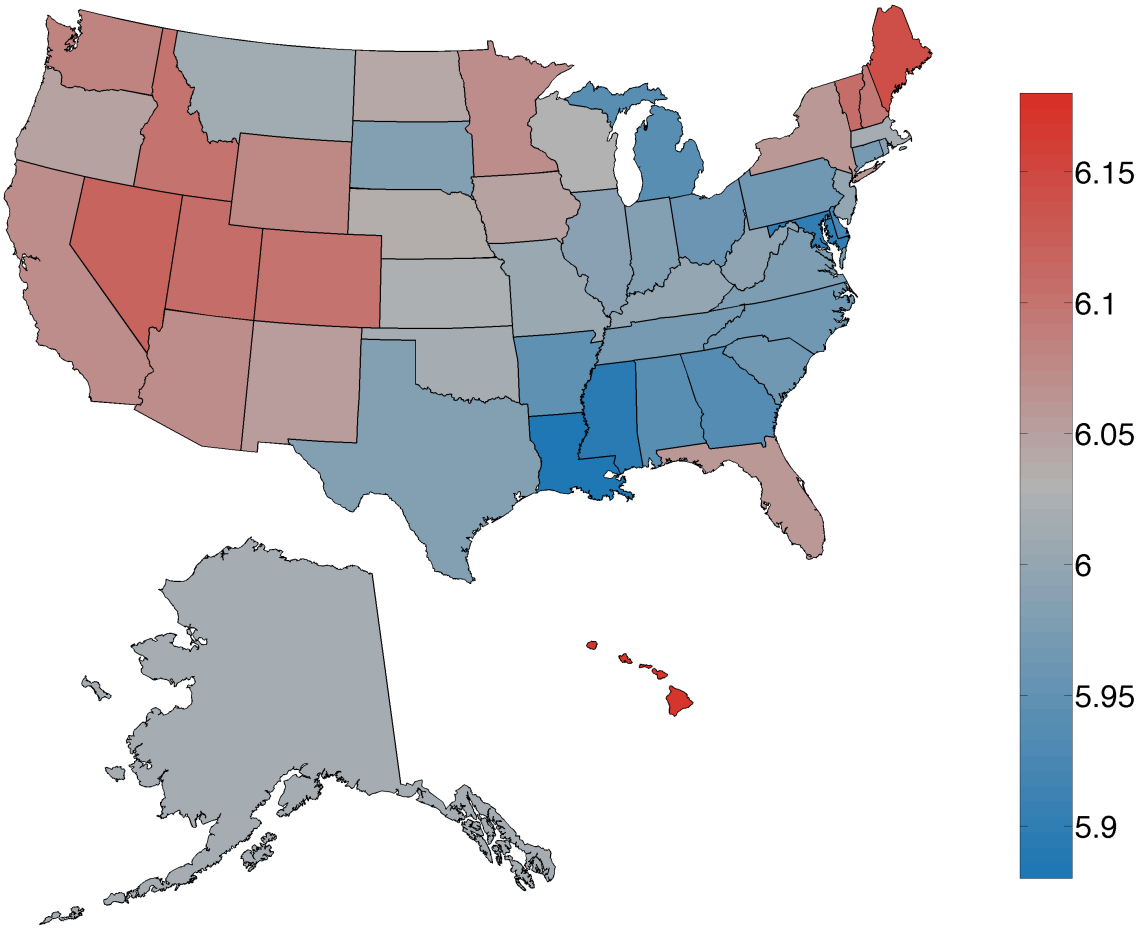

FIG. 1: Choropleth showing average word happiness for geotagged tweets in all US states collected during the calendar year 2011. The happiest 5 states, in order, are: Hawaii, Maine, Nevada, Utah and Vermont. The saddest 5 states, in order, are: Louisiana, Mississippi, Maryland, Delaware and Georgia. Word shift plots describing how differences in word usage contribute to variation in happiness between states are presented in Appendix B (online).

to the other states, in stark contrast with the findings of Oswald and Wu [25] that Louisiana exhibited the highest score on an alternate measure of life satisfaction.

We can further use this data on word frequencies to characterize similarities between states based on word usage. Figure 2 shows the linear correlation between word frequency vectors $\mathbf{f} = \{f_i, i = 1 : 50000\}$ for each pair of states, with red entries in the matrix indicating states with similar word use. We see some clusters which might be explained by geographical proximity, such as Vermont and New Hampshire or Louisiana and Mississippi, and some outliers such as the state of Nevada, which correlates the lowest on average with all other states. Additional details on this state-level dataset, including plots of raw number of tweets and number of tweets per head of population for each state can be found in Appendix A. Word shift graphs showing which words contribute most to the variation in happiness across states can be found in Appendix B (online) [1].

We now change our resolution to a finer scale by focussing on cities rather than states. As an illustration of the resolution of the data set as well as our technique, we plot a tweet-generated map of a city, showing how average word happiness varies with location. In figure 3 we plot tweets collected from the New York City area during 2011. Each point represents an individual tweet, and is colored by the happiness h_{avg} of the text T consisting of the $N = 200$ closest LabMT words to the location of that tweet. We set a maximum threshold radius of $r = 500$ meters around each tweet location; if 200 LabMT words cannot be found within that radius then the point is colored black. Several features can immediately be discerned in this purely tweet-generated map. Firstly, the spatial resolution reveals the outline of Manhattan, as well as Central Park, individual streets and bridges, and even airport terminals such as those at JFK and Newark airports at the lower right and center left of the figure respectively. Secondly, we can discern

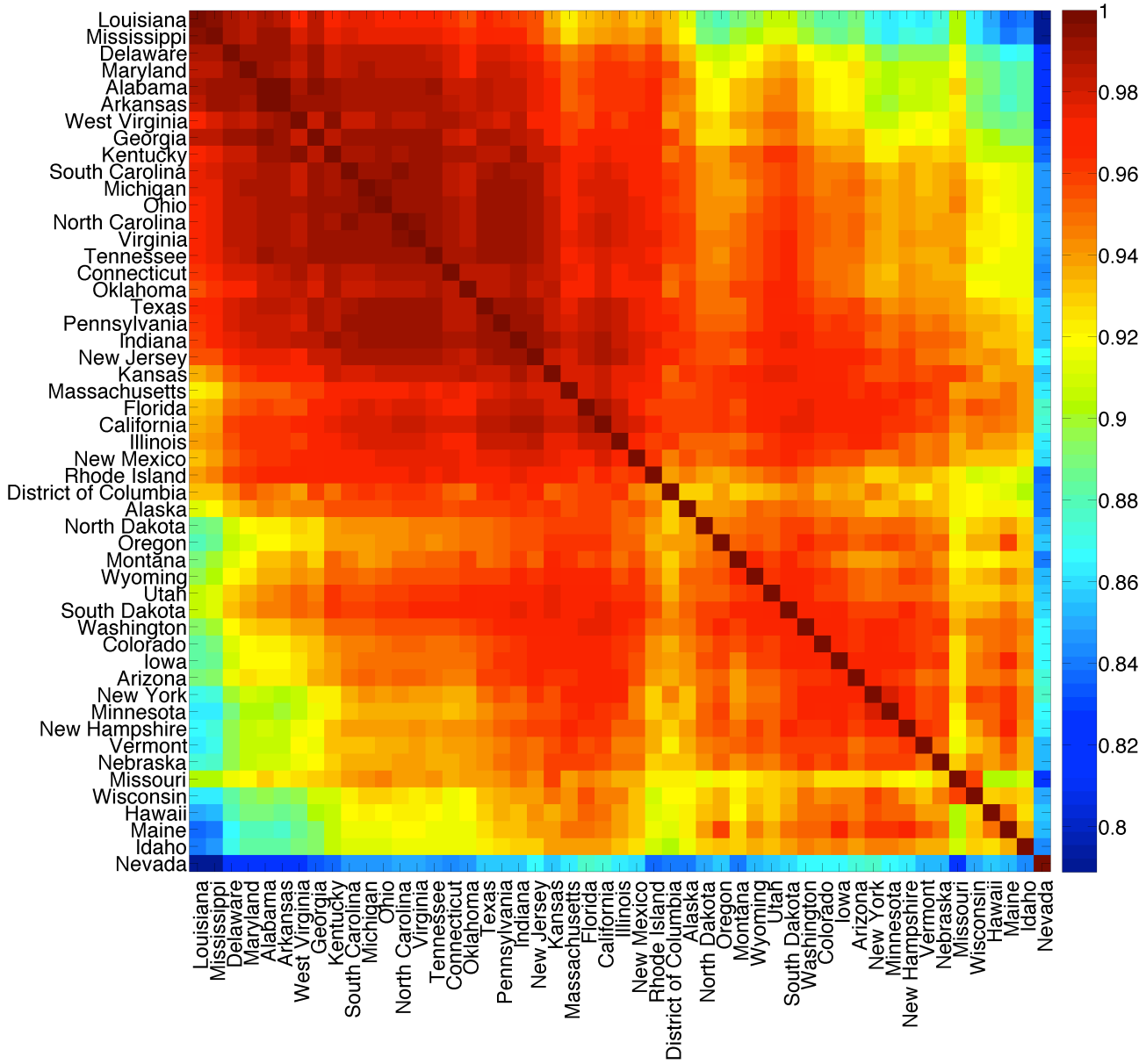

FIG. 2: Clustergram showing cross-correlations between word frequency distributions for all states in 2011. Red signifies states with similar or highly-correlating word frequency distributions, while blue signifies states with relatively dissimilar word frequency distributions.

regions of higher and lower happiness: the Harlem and Washington Heights areas to the north appear relatively sad compared to the Downtown/Midtown area, as does the Waterfront, New Jersey area west of the southern tip of Manhattan. Similar tweet-generated maps for all 373 cities in the data set are presented in Appendix B (online) [1].

In figure 4 we show a tweet-generated happiness map of the entire contiguous United States, where we have now used $N = 500$ and $r = 10$ km. We can clearly discern cities and the roads between them at this scale, and substantial variation in happiness across geographical regions. There is already an indication that some cities will be significantly less happy than others, particularly those in the southeastern United States, a conclusion which will be made more quantitative later. At a finer scale we can see that some coastal areas, particularly around the Florida peninsula and along the coast of

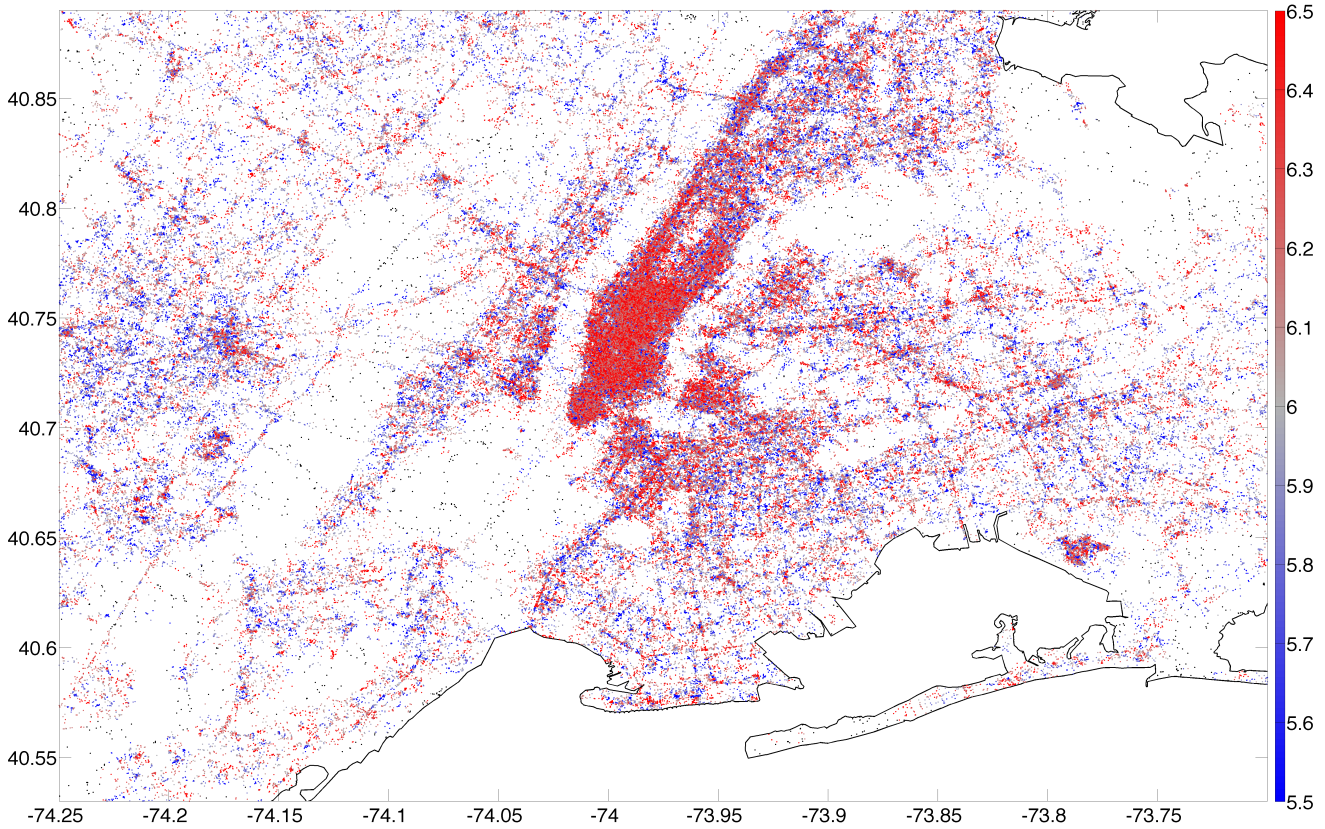

FIG. 3: Map of tweets collected from New York City during the calendar year 2011. Each point represents an individual tweet and is colored by the average word happiness h_{avg} of nearby tweets: red is happier, blue is sadder. For a point to be colored, we require that there be at least 100 LabMT words within a 500 meter radius of the location; points which do not satisfy this criterion are colored black.

North and South Carolina, are significantly happier than the regions immediately inland of them. We will see this again below in the word shifts for various beachside cities.

Next we calculate the happiness h_{avg} for each city in the census data set using equation (1), where the boundaries of a city are defined by the MAF/TIGER database, and each text T is formed by agglomerating all the words falling within that city. Figure 5 shows the distribution of happiness scores for all cities; as is to be expected for smaller samples, the range of values is slightly higher than that calculated for the states, extending over a range of more than 0.2 from the mean of $h_{\text{avg}} = 6.00$. We remark that the distributed is skewed: there are more cities that are happier than the overall average, by 220 to 153.

It is well known that city population sizes follow a power law distribution (see [36] and many others), which in conjunction with figure 5 suggests that happiness decreases with city size. While we did find a slight negative correlation between happiness and the number of tweets gathered in each city, we in fact found that happiness strongly negatively correlates with the number of tweets per capita, with Spearman correlation coefficient -0.558 and p-value less than 10^{-16}, as shown in figure 6. This suggests that cities with high technology adoption rates (as most geotagged tweets come from devices like smartphones) are in fact less happy than their less-technological counterparts.

The bar charts in figures 7 and 8 show the average word happiness h_{avg} for the 15 happiest and 15 saddest cities in the contiguous United States, respectively. Using this method we identify Napa, California as the happiest city in the US with a score of 6.26, and Beaumont, Texas as the saddest city with a score of 5.83.

Perhaps surprisingly, several cities that ranked both highly and lowly by our measure rank similarly in more traditional survey based efforts. For example, a Gallup-Healthways well-being survey for 2011 [19] showed Boulder, Colorado as the city with the fifth highest well-being index composite score (and twelfth highest happiness score in our list), while Flint, Michigan had the second lowest and Montgomery, Alabama the 21st-lowest well-being index (compared to 8th lowest and 14th lowest happiness scores on our list). The overall Spearman correlation between the rankings using Gallup's well-being index and with our measure is $\rho = 0.328$, with p-value 7.73×10^{-6} (a scatter plot is presented online in Appendix C). Whereas our list uses only word frequencies in the

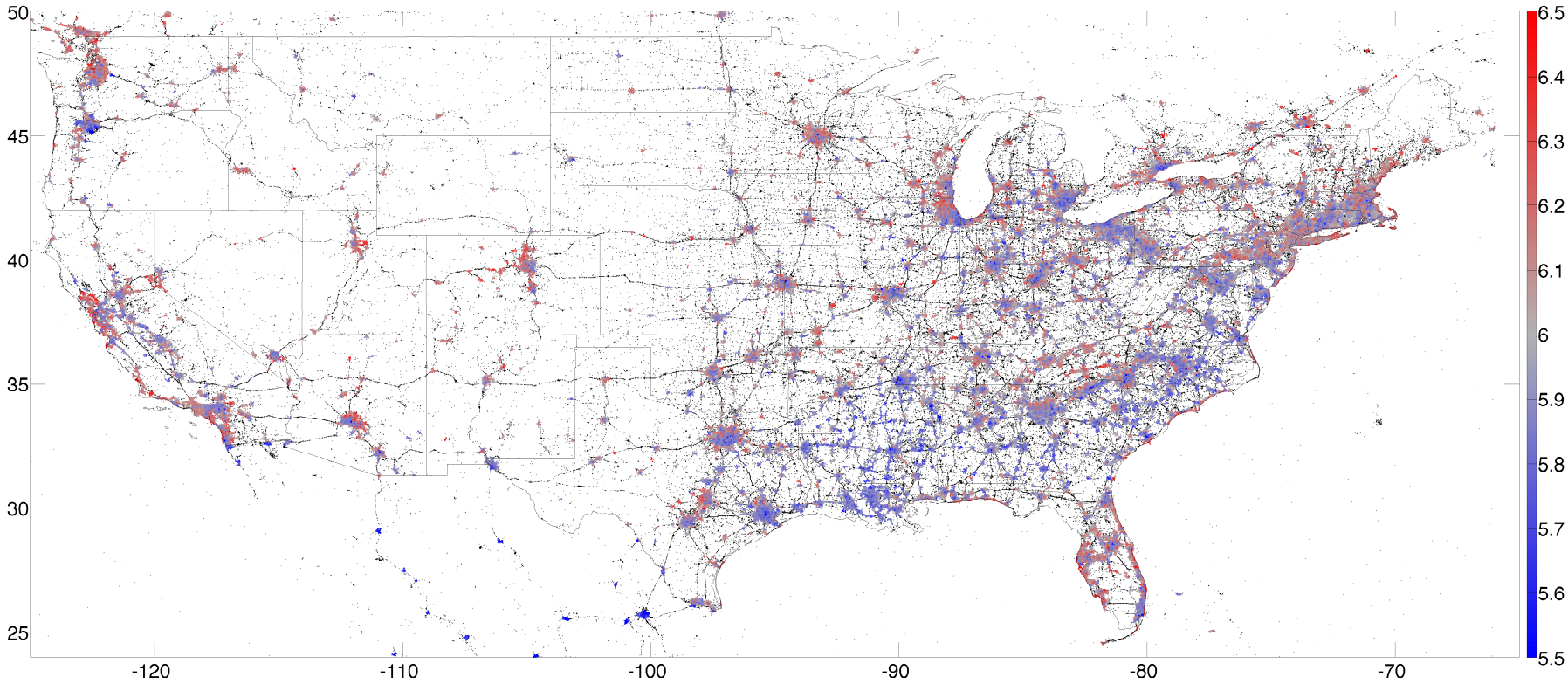

FIG. 4: Map showing happiness of all tweets collected from the lower 48 states during the calendar year 2011. Points are colored as in figure 3, except we now require that there are at least 500 LabMT words within a 10 kilometer radius of the location of each tweet in order to be colored.

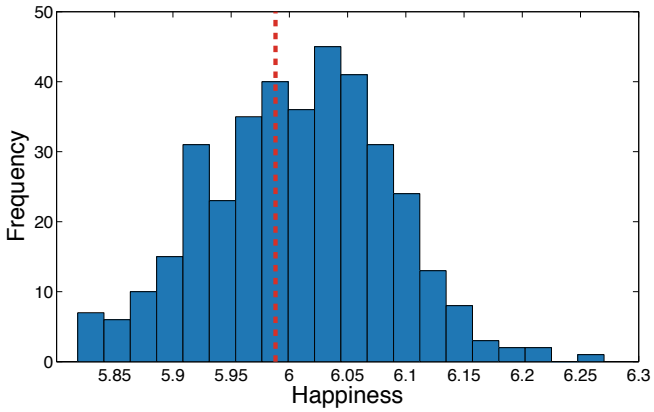

FIG. 5: Histogram showing the distribution of happiness values for the 373 cities in the census data set. A vertical dashed line denotes the average for all cities. Note the greater weight towards the right of the distribution, with more cities having happiness scores higher than the average.

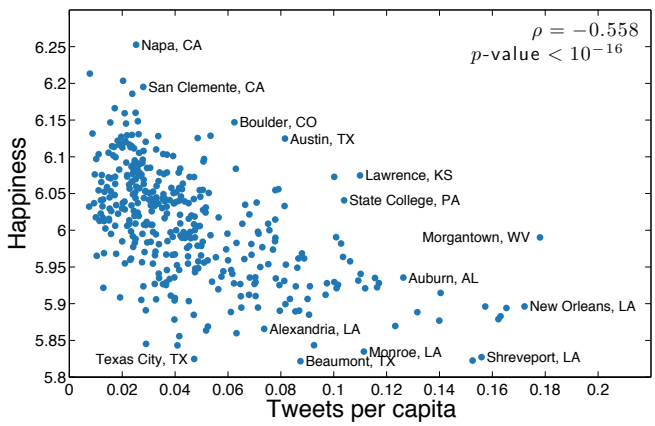

FIG. 6: Happiness as a function of number of tweets per capita. Areas with a higher density of tweets per capita tend to be less happy.

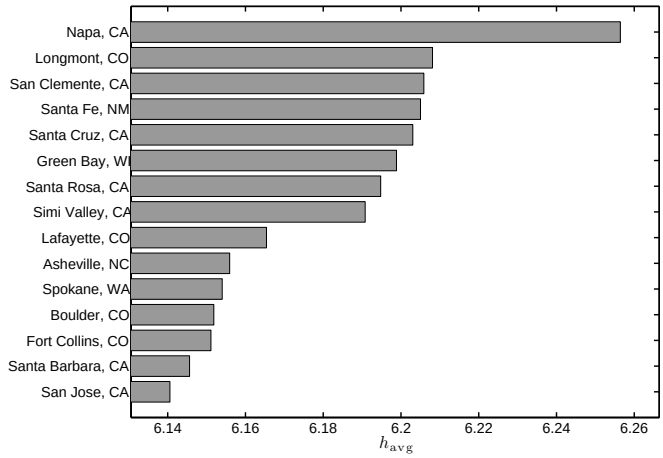

FIG. 7: The 15 highest average word happiness scores $h_{\rm avg}$ for cities in the contiguous USA, as calculated using (1) and the LabMT word list.

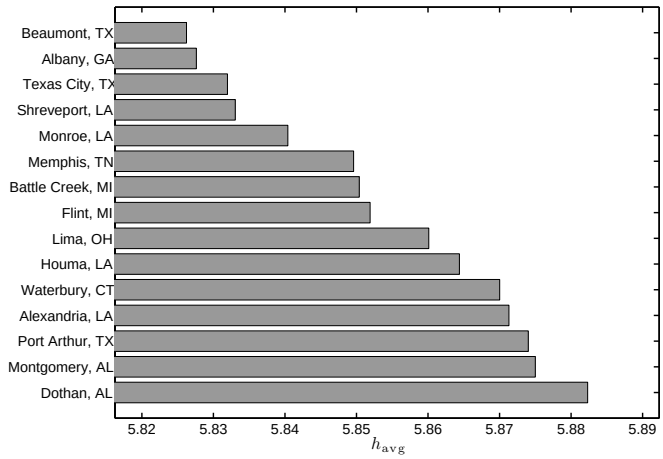

FIG. 8: The 15 lowest average word happiness scores $h_{\rm avg}$ for cities in the contiguous USA, as calculated using (1) and the LabMT word list.

calculation of $h_{\rm avg}$, the Gallup-Healthways score is an average of six indices which measure life evaluation, emotional health, work environment, physical health, healthy behaviors, and access to basic necessities. We remark that our method is (a) far more efficient to implement than a survey-based approach, and (b) provides a near real-time stream of information quantifying well-being in cities.

To investigate why the average word happiness varies across urban areas, we study the word shift graphs [10, 11] for each city. These graphs show how the difference in happiness for two texts depends on differences in the underlying word frequencies. In figure 9 we show the word shift graphs for Napa and Beaumont, as compared to the entire corpus of words collected for all urban areas during 2011. Word shift graphs for every city are presented in Appendix C (online) [1].

We observe some features of the graphs that are consistent with geography—for example the word 'beach' appears high on the list of words for coastal cities such as Santa Cruz, California or Miami, Florida. Overall, the main factor driving the relative happiness scores for each city appears to be the presence or absence of key words such as 'lol', 'haha' and its variants, 'hell', 'love', 'like', as well as profanity.

IV. CORRELATING WORD USAGE WITH CENSUS DATA

The word shifts of figure 9 demonstrate how word usage varies with location, as well as the importance of studying the individual words that go in to the calculation of averaged quantities such as the word happiness

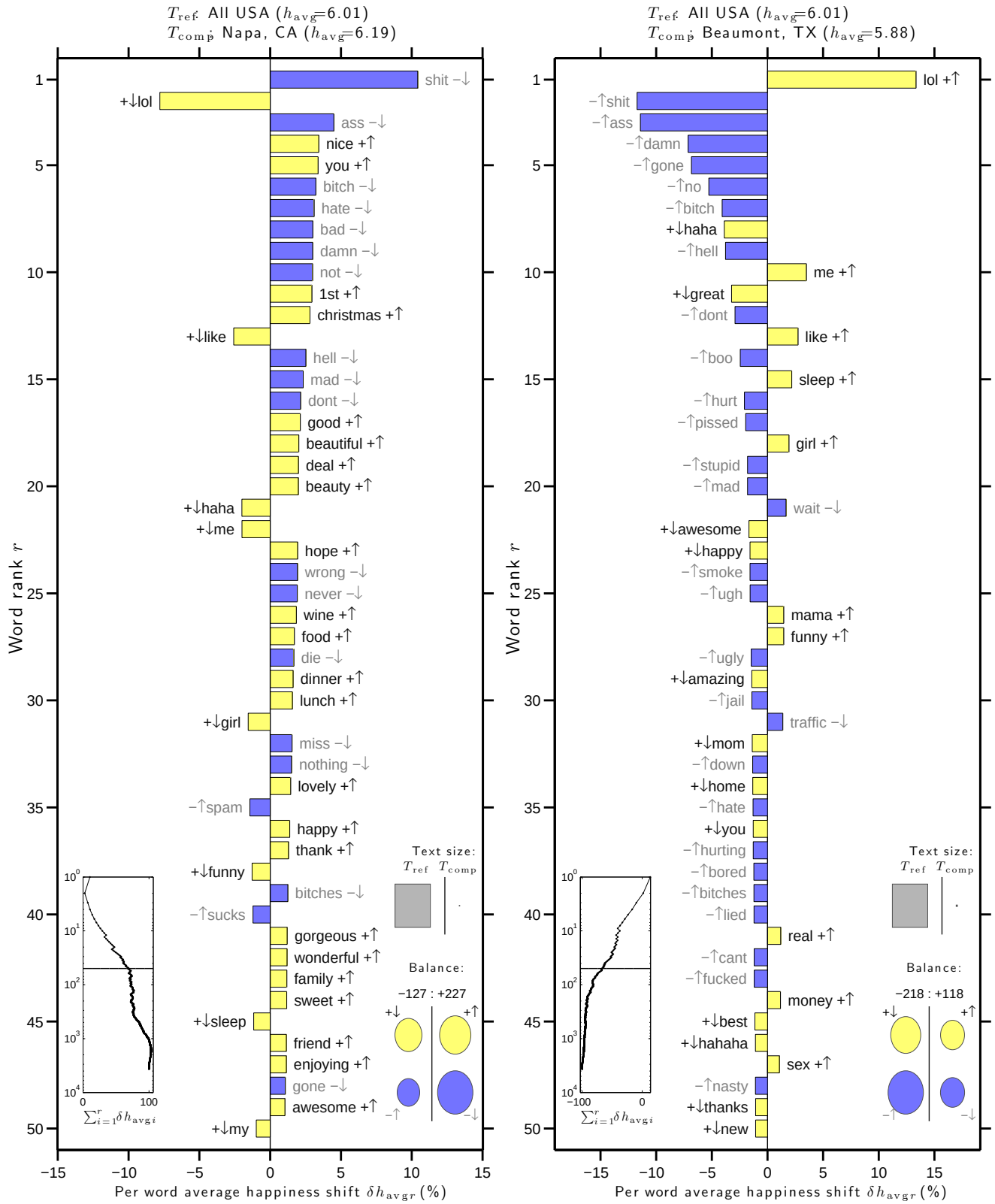

FIG. 9: Word shift graphs showing how $h_{\rm avg}$ varies for all US cities measured versus the cities Napa, California (left) and Beaumont, Texas (right) with highest and lowest $h_{\rm avg}$ respectively. Words are ranked in order of decreasing percentage contribution to the overall average happiness difference $\delta h_{\rm avg}$. The symbols $+/-$ indicate whether a word is relatively happy or sad compared to $h_{\rm avg}$ for the entire US (text $T_{\rm ref}$), while the arrows $\uparrow$ / $\downarrow$ indicate whether the word was used more or less in the text $T_{\rm comp}$ for each city than in $T_{\rm ref}$. The left inset panel shows how the ranked LabMT words combine in sum. The four circles at bottom right show the total contribution of the four kinds of words ($+\downarrow$, $+\uparrow$, $-\uparrow$, $-\downarrow$). Relative text size is indicated by the areas of the gray squares.

h_{avg}. We now therefore examine in greater detail how happiness and word usage relate to underlying social factors.

We first focus on how the average happiness h_{avg}, correlates with different social and economic measures. To do this we took data from the American Community Survey 1-year estimates for 2011, specifically tables DP02 through DP05, covering selected social characteristics, economic characteristics, housing characteristics and demographic and housing estimates. These tables contained 508 different categories for all cities, from which we removed the categories with data on less than 75% of all cities, leaving 432 different categories for correlation with happiness.

In figure 10 we show the Spearman correlation between happiness and each demographic attribute for cities in the census data set. Each point in the graph represents one of the 432 attributes considered; a table listing each demographic and its correlation with happiness is presented in Appendix D (online) [1]. The groupings into columns were made independently of happiness values, by performing distance-based clustering using a hierarchical cluster tree on the table of census attributes for all cities. The 8 clusters which were found are not unique and depend on the distance threshold used, however they give some indication of which attributes covary. Only two groups show a large number of attributes which significantly correlate (below $p = 0.01$) with happiness; these are shown in blue (with red crosses specifying the median attribute). These two groups might be broadly characterized as representing high socioeconomic and low socioeconomic status respectively, with many of the attributes in the high socioeconomic status group positively correlating with happiness (and vice versa for the low socioeconomic status group).

To further understand what drives this correlation of certain demographics with happiness, we now investigate how each word from the LabMT list correlates with all attributes from the census. To do this we first normalize the word counts in each urban area by the total number of tweets collected in each city, and then for each word calculate the Spearman correlation ρ between normalized frequency and census attribute for all cities. For example, the scatter plot in figure 11 shows that the normalized frequency of occurrence of the word 'cafe' shows a strong positive correlation with the percentage of the population with a bachelor's degree or higher. The Spearman correlation between the two is $\rho = 0.481$ with p-value 4.90×10^{-23}, indicating strong correlation.

Lists showing the correlation of each LabMT word with every demographic attribute are presented in Appendix D (online) [1]. Taking the percentage of population with a bachelors degree or higher for urban areas from the 2011 census as a representative example, tables I and II show the top 25 words which show the highest positive and negative correlations respectively to this attribute. The results show that longer words such as 'software', 'development' and 'emails' correlate strongly with education, while the words which correlate negatively with education are generally shorter, with no words longer than two syllables appearing in the list. Furthermore, many of the words such as 'love', 'talk' and 'mom' appearing in table II are family- or relationship-oriented, while the more technical terms appearing in table I are more employment-oriented, and suggest more complex and abstract intellectual themes. It may be postulated that this is a reflection of the social processes occurring in urban areas characterized by rates of low and high education, respectively.

Word	ρ	p-value	$h_{\text{avg}}(w_i)$
cafe	0.481	4.9×10^{-23}	6.78
pub	0.463	3.14×10^{-21}	6.02
software	0.458	9.07×10^{-21}	6.30
yoga	0.455	1.85×10^{-20}	7.04
grill	0.433	1.78×10^{-18}	6.24
development	0.424	1.14×10^{-17}	6.38
emails	0.419	2.87×10^{-17}	6.54
wine	0.417	3.83×10^{-17}	6.42
library	0.414	6.47×10^{-17}	6.48
art	0.414	6.8×10^{-17}	6.60
sciences	0.410	1.54×10^{-16}	6.30
pasta	0.410	1.57×10^{-16}	6.86
lounge	0.409	1.68×10^{-16}	6.50
market	0.408	2.2×10^{-16}	6.28
india	0.407	2.5×10^{-16}	6.42
drinking	0.405	3.74×10^{-16}	6.14
technology	0.405	3.76×10^{-16}	6.74
forest	0.405	3.83×10^{-16}	6.68
brunch	0.405	3.89×10^{-16}	6.32
dining	0.403	4.92×10^{-16}	6.48
supporting	0.399	1.1×10^{-15}	6.48
professor	0.398	1.23×10^{-15}	6.04
university	0.392	3.62×10^{-15}	6.74
film	0.391	4.27×10^{-15}	6.56
global	0.391	4.72×10^{-15}	6.00

TABLE I: Top 25 words with strongest positive Spearman correlation ρ to percentage of population with a Bachelors degree or higher (census table DP02-HC03-VC94) in 2011. Stop words with $4 < h_{\text{avg}} < 6$ have been removed from the list. Note the low p-values for all words, indicating strong statistical significance.

The technique applied here is not limited to only the traditional types of data collected through the census. As an example of a different use of use of the data set, we correlate word use to obesity at the metropolitan level. For this study we take obesity levels from the Gallup and Healthways 2011 survey [35], and metropolitan areas as defined by the U.S. Office of Management and Budget's definitions for Metropolitan Statistical Areas (MSAs) [31]. We remark that the MSAs are generally two to three

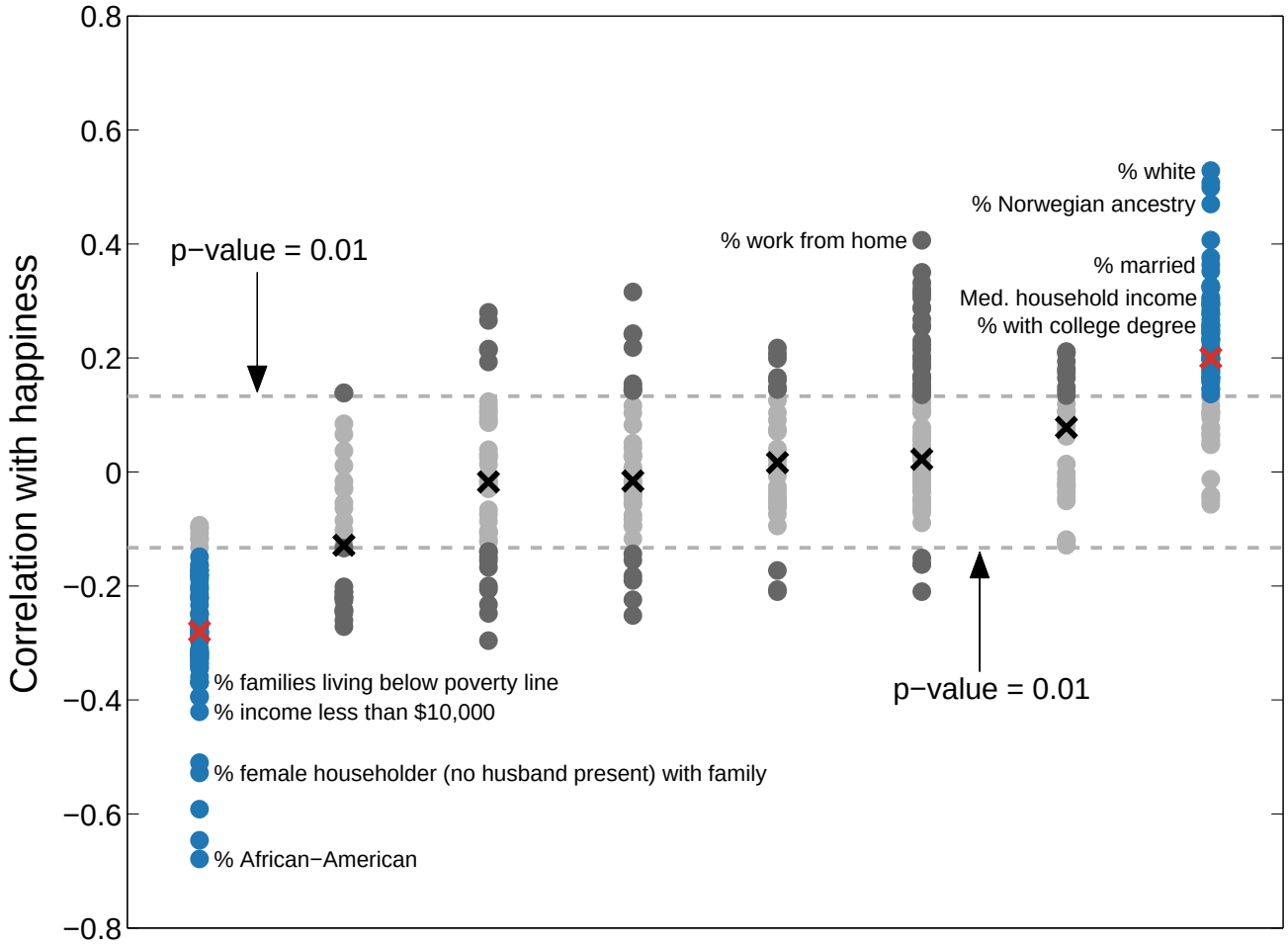

FIG. 10: Spearman correlations for 432 demographic attributes with happiness. The 8 groupings along the horizontal axis are for covarying attributes identified by agglomerative hierarchical clustering, independently of happiness. Crosses lie on the median of each cluster, and the dashed lines represent the 1% significance level. The two clusters which have medians that correlate significantly with happiness are colored blue. A complete list of the correlation of all attributes with happiness can be found in Appendix D (online).

times larger in area than the TIGER urban area census boundaries, and the Gallup obesity survey was only for the 190 largest-population areas. The obesity data set contains fewer small cities than the TIGER census set, particularly in the midwest. We collected more than 10 million tweets from these 190 MSAs, corresponding to just over 80 million words during 2011.

Performing the same analysis as for the attributes in figure 10, in figure 12 we show the relationship between happiness and obesity for the 190 MSAs included in the Gallup survey. We find that happiness generally decreases as obesity increases, with the third happiest city in this set (Boulder, Colorado) corresponding with the lowest obesity rate (12.1%) and the saddest city (Beaumont, Texas, as found previously) corresponding with the fifth highest obesity rate (33.8%). We calculate a Spearman correlation coefficient of $\rho = -0.339$ with p-value 2.01×10^{-6} for the data, indicating statistically significant negative correlation.

As previously for the census data, we also correlate the abundance of each individual word in the LabMT list to obesity levels in the 190 cities surveyed. From this list we extract words that are clearly food-related, and present those which most most strongly negatively and positively correlate with obesity in table III. Note that we are including stop words for which $4 < h_{\rm avg}(w_i) < 6$ in these lists. Coffee-related words such as 'cafe', 'coffee', 'espresso' and 'bean' feature prominently in the list, and many of the words refer to eating at restaurants—'sushi', 'restaurant', 'cuisine' and 'brunch', for example. As we might expect such words to correlate with wealth, this suggests a correlation between obesity and poverty, a claim which we note remains contentious in the medical literature (for example, supported in [21, 24], and refuted in [7]).

Conversely, only 6 food-related words significantly pos-

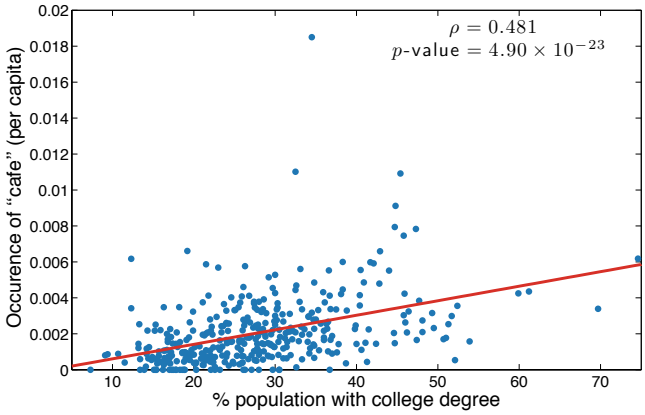

FIG. 11: Scatter plot showing the correlation between rate of occurrence of the word 'cafe' and percentage of population with a bachelor's degree or higher in US cities during the calendar year 2011. The red line shows linear correlation while the reported ρ and p-values show the Spearman correlation.

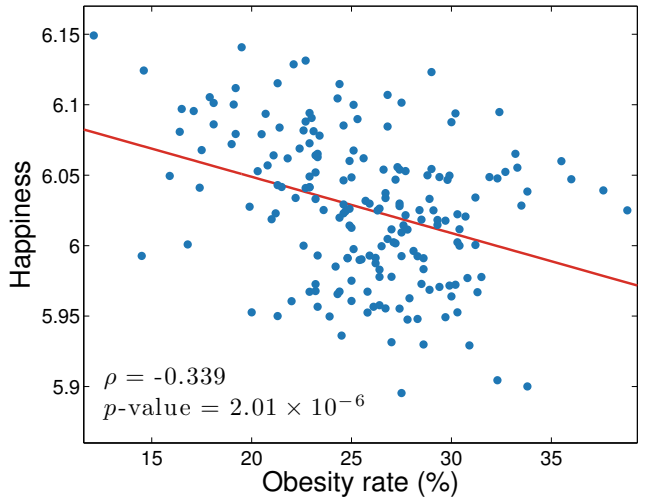

FIG. 12: Scatter plot showing correlation between h_{avg} and obesity level, as taken from the 2011 Gallup and Healthways survey. The red line is the straight line of best fit to the data, while the ρ value is the Spearman correlation coefficient for the data.

itively correlate with obesity with p-values less than 0.05 (note again the asymmetry in the number of words which positively and negatively correlate with obesity). The fast food chain 'mcdonalds' correlates most strongly, and the foods 'wings' and 'ham' both appear. Unlike in the low-obesity word table, words describing a desire for food—'eat' and 'hungry'—as well as the negative reaction of 'heartburn' to overeating, both appear on the list. In Appendix A we show tables listing the food-related words which show the least correlation with obesity, as well as the top 25 words (food-related or not) from the LabMT list that correlate and anti-correlate with obesity. The full list of LabMT words and their correlations with obesity can be found in Appendix E (online) [1].

Word	ρ	p-value	$h_{\text{avg}}(w_i)$
me	-0.393	3.26×10^{-15}	6.58
love	-0.389	6.51×10^{-15}	8.42
my	-0.354	1.97×10^{-12}	6.16
like	-0.346	6.04×10^{-12}	7.22
hate	-0.344	8.76×10^{-12}	2.34
tired	-0.343	1×10^{-11}	3.34
sleep	-0.341	1.27×10^{-11}	7.16
stupid	-0.328	8.55×10^{-11}	2.68
bored	-0.315	5.11×10^{-10}	3.04
you	-0.315	5.23×10^{-10}	6.24
goodnight	-0.305	1.77×10^{-9}	6.58
bitch	-0.295	6.51×10^{-9}	3.14
all	-0.289	1.33×10^{-8}	6.22
lie	-0.285	2.24×10^{-8}	2.60
mom	-0.284	2.42×10^{-8}	7.64
wish	-0.271	1.05×10^{-7}	6.92
talk	-0.267	1.74×10^{-7}	6.06
she	-0.265	2.01×10^{-7}	6.18
know	-0.262	2.78×10^{-7}	6.10
ill	-0.259	4.11×10^{-7}	2.42
dont	-0.258	4.54×10^{-7}	3.70
well	-0.256	5.3×10^{-7}	6.68
don't	-0.255	5.8×10^{-7}	3.70
give	-0.255	5.84×10^{-7}	6.54
friend	-0.255	6.27×10^{-7}	7.66

TABLE II: Top 25 words with strongest negative Spearman correlation ρ to percentage of population with a Bachelors degree or higher in 2011 (with stop words removed).

The above analysis demonstrates that different cities have unique characteristics. We now ask whether cities can be sorted into groups solely based upon similarities in their word distributions. Bettencourt *et al.* [6] used data on the economy, crime and innovation to characterize cities; here we use a similar methodology except with word frequency data to uncover so-called 'kindred' cities.

We group the top 40 cities with highest word counts in 2011 by calculating the linear correlation between word frequency vectors $\mathbf{f}$ as we did in Figure 2. The resulting cross-correlation matrix is shown in figure 13, with red signifying strong correlation between cities. Firstly we note that all cities show similar word frequency distributions, with all correlations being higher than $\rho = 0.8$. As was the case for the states (see figure 2), we see one clear large group of strongly correlated cities emerge in the lower right corner, with a smaller distinct cluster appearing at the top left. Perhaps uniquely, these groupings are defined solely by similarities in word usage between cities, rather than by geography or economic indicators.

We cluster cities using an agglomerative hierarchical method with average linkage clustering, as shown in the

Word	ρ	p-value	$h_{\mathrm{avg}}(w_i)$
cafe	-0.509	6.07×10^{-14}	6.78
sushi	-0.487	9.93×10^{-13}	5.40
brewery	-0.469	8.67×10^{-12}	N/A
restaurant	-0.448	8.93×10^{-11}	7.06
bar	-0.435	3.59×10^{-10}	5.82
banana	-0.434	3.77×10^{-10}	6.86
apple	-0.408	5.22×10^{-9}	7.44
fondue	-0.403	8.34×10^{-9}	N/A
wine	-0.400	1.08×10^{-8}	6.42
delicious	-0.392	2.17×10^{-8}	7.92
dinner	-0.386	3.85×10^{-8}	7.40
coffee	-0.384	4.51×10^{-8}	7.18
bakery	-0.383	5.12×10^{-8}	N/A
bean	-0.378	7.88×10^{-8}	5.80
espresso	-0.377	8.47×10^{-8}	N/A
cuisine	-0.376	8.82×10^{-8}	N/A
foods	-0.374	1.07×10^{-7}	7.26
tofu	-0.372	1.27×10^{-7}	N/A
brunch	-0.368	1.79×10^{-7}	6.32
veggie	-0.364	2.46×10^{-7}	N/A
organic	-0.361	3.13×10^{-7}	6.32
booze	-0.360	3.34×10^{-7}	N/A
grill	-0.354	5.4×10^{-7}	6.24
chocolate	-0.351	6.77×10^{-7}	7.86
#vegan	-0.350	7.47×10^{-7}	N/A
mcdonalds	0.246	6.18×10^{-4}	5.98
eat	0.241	8.22×10^{-4}	7.04
wings	0.222	2.13×10^{-3}	6.52
hungry	0.210	3.65×10^{-3}	3.38
heartburn	0.194	7.37×10^{-3}	N/A
ham	0.177	1.45×10^{-2}	5.66

TABLE III: The top 25 food-related words only with strongest negative correlation to obesity level (top), and the 6 food-related words with positive correlation to obesity level and p-value less than 0.05 (bottom).

dendrogram at the top of figure 13, and highlight the 4 clusters with lowest linkage threshold using different colors. As one might expect, some cities that are geographically nearby are grouped together. Notable examples are some cities in the southern US such as Baton Rouge, New Orleans and Memphis in the lower right of the plot, as well as the Californian cities of San Diego and San Francisco at top left. However, this pattern does not hold for all cities; while there is the suggestion of a north/south grouping between the two clusters at the top left and the two at the bottom right, some cities such as Austin and Tampa in the south and Detroit and Philadelphia in the north go against this trend. The cities of Cleveland and Detroit are the most alike in word use, having a cross-correlation of $\rho = 0.995$, while Austin and Baton Rouge are the most dissimilar with a cross-correlation of $\rho = 0.813$. Indianapolis is the city with highest average correlation to the word use in other cities ($\bar{\rho} = 0.961$), while Minneapolis shows the most unique word use on average, with $\bar{\rho} = 0.884$.

V. DISCUSSION

In this paper we have examined word use in urban areas in the United States, using a simple mathematical method which has been shown to have great flexibility, sensitivity and robustness. We have used this tool to map areas of high and low happiness and score individual cities for average word happiness. In order to understand in greater detail how word usage influences happiness, we used both word shift graphs to find the words which produced the most difference between the happiness scores of each city and the average for the entire US, and socioeconomic census data to attempt to explain the usage of certain words. A significant driver of the happiness score for individual cities was found to be frequency of swear word use; we believe that future studies of regional variation in swear word use or 'geoprofanity' could help explain geographical differences in happiness. Indeed, swearing has previously been found to be a predictor of large-scale protests and social uprisings in Iran [15].

Happiness within the US was found to correlate strongly with wealth, showing largest positive correlation with household income and strongest negative correlation with poverty amongst the census data sets used. This is consistent with the first part of the 'Easterlin paradox' [13], that within countries at a given time happiness consistently increases with income. The second part of the paradox is that while personal wealth has been observed to consistently increase over time, happiness has tended to decrease in both developed and developing countries [13, 14]. A previous result using this method showing a decline in happiness over the 2009-2011 period (see figure 3 of [11]) is consistent with this finding. The relationship between wealth and happiness is still highly debated; recent works by Stevenson and Wolfers [32] claim to show a direct correlation between gross domestic product and subjective well-being across countries, while Di Tella and MacCulloch [8] in the same year argue that the Easterlin paradox is in fact exacerbated if other economic variables than just income are considered.

Interestingly, happiness was also observed to anticorrelate significantly with obesity. A similar link between obesity and happiness has previously been reported [17], particularly for individuals who report low self control [33]. However, as some authors point out, the presence of chronic illnesses accompanying obesity can confound the link between obesity and psychological well-being [12], and indeed an inverse relationship between weight and depression has been found in some studies [26]. We remark that it should be possible to use techniques such as those described here to mine social network data for

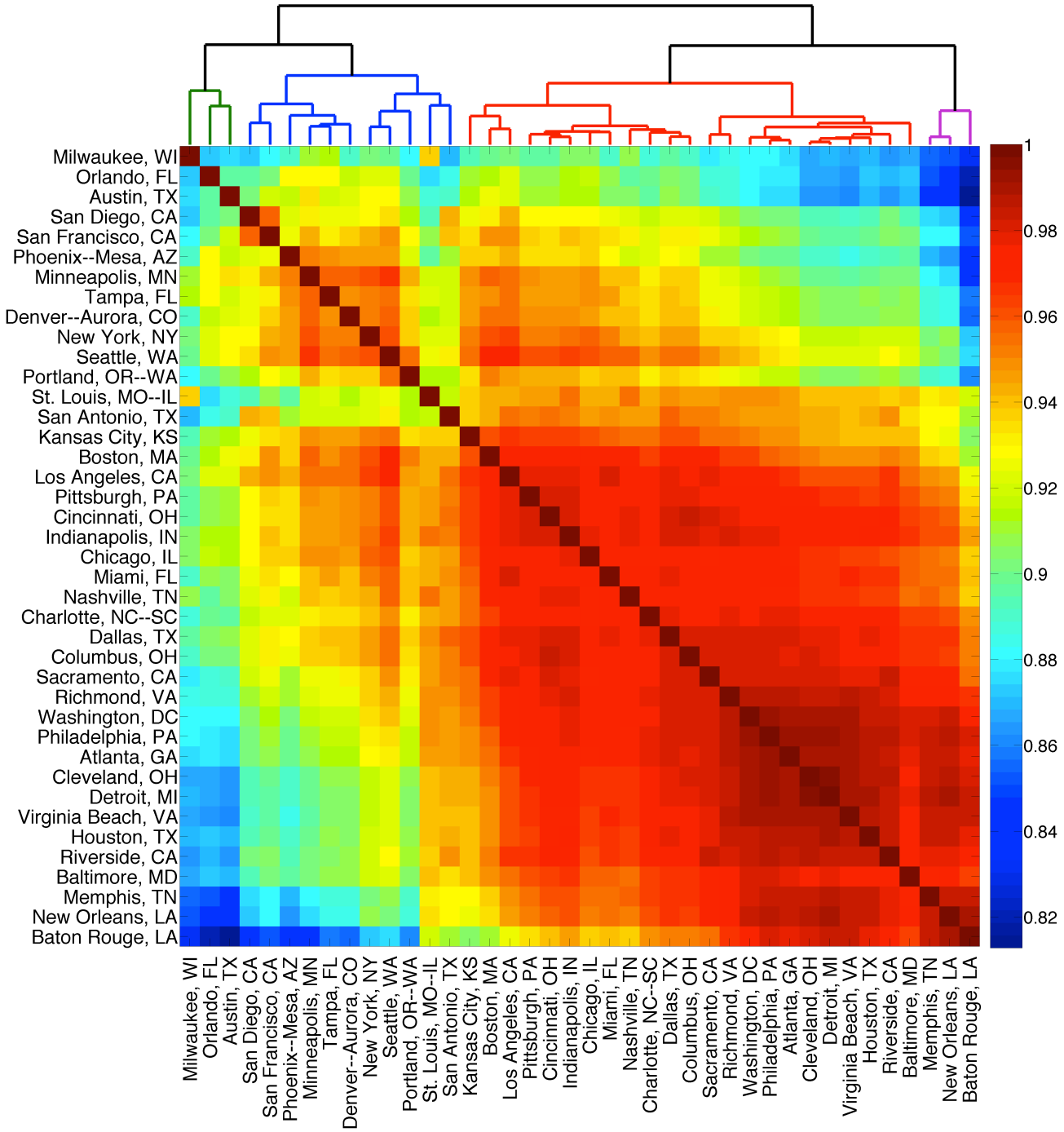

FIG. 13: Cross-correlations between word frequency distribution differences for the 40 cities with highest word counts. Red signifies cities with similar word frequency distribution, while blue signifies cities with dissimilar word frequency distributions.

real-time surveying. For example, the potential for identifying areas with high obesity based solely on word use is significant.

There are a number of legitimate concerns to be raised about how well the Twitter data set can be said to represent the happiness of the greater population. Only 15% of online adults regularly use Twitter, and 18-29 year-olds and minorities tend to be more highly represented on Twitter than in the general population [30]. Furthermore, the fact that we collected only around 10% of all tweets during the calendar year 2011 means that our data set is a non-uniform subsample of statements made by a

non-representative portion of the population.

In this work we have only scratched the surface of what is possible using this particular dataset. In particular, we have not examined whether or not these methods have any predictive power—future research could look at how observed changes in the Twitter data set, as measured using the hedonometer algorithm, predict changes in the underlying social and economic characteristics measured using traditional census methods. In particular, we plan to revisit this study when census data for 2012 becomes available to investigate how changes in demographics across urban areas is reflected in happiness as measured by word use.

Acknowledgments

The authors are grateful for the computational resources provided by the Vermont Advanced Computing Core which is supported by NASA (NNX 08A096G), and the Vermont Complex Systems Center. CMD and LM were supported by NSF grant DMS-0940271 and PSD was supported by NSF CAREER Award #0846668. The authors also wish to acknowledge support from the MITRE Corporation.

[1] Supplementary material for this article is avail-able online at http://www.uvm.edu/storylab/share/papers/mitchell2013a/.

[2] Amazon Mechanical Turk. https://www.mturk.com/, accessed February 2013.

[3] Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/, accessed Februrary 2013.

[4] Jen Beaumont and Jennifer Thomas. Measuring National Well-being - Health. Technical Report July, 2012.

[5] Luís M. A. Bettencourt, J. Lobo, Dirk Helbing, Christian Kuhnert, and Geoffrey B. West. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences*, 104(17):7301–7306, 2007.

[6] Luís M. A. Bettencourt, José Lobo, Deborah Strumsky, and Geoffrey B. West. Urban scaling and its deviations: revealing the structure of wealth, innovation and crime across cities. *PloS one*, 5(11):e13541, January 2010.

[7] Virginia W. Chang and Diane S. Lauderdale. Income disparities in body mass index and obesity in the United States, 1971-2002. *Archives of Internal Medicine*, 165(18):2122–8, October 2005.

[8] R. Di Tella and R. MacCulloch. Gross national happiness as an answer to the Easterlin Paradox? *Journal of Development Economics*, 44(617), 2008.

[9] US Census Bureau Geography Division. 2010 Census TIGER/Line Shapefiles. https://www2.census.gov/geo/pdfs/maps-data/data/tiger/tgrshp2010/TGRSHP10SF1.pdf, accessed February 2013.

[10] Peter Sheridan Dodds and Christopher M. Danforth. Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents. *Journal of Happiness Studies*, 11(4):441–456, July 2009.

[11] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PloS one*, 6, 2011.

[12] H. A. Doll, S. E. Petersen, and S. L. Stewart-Brown. Obesity and physical and emotional well-being: associations between body mass index, chronic illness, and the physical and mental components of the SF-36 questionnaire. *Obesity Research*, 8(2):160–70, March 2000.

[13] Richard A. Easterlin. Does economic growth improve the human lot? Some empirical evidence. *Journal of Economic Behavior and Organization*, 27:35–47, 1974.

[14] Richard A. Easterlin, Laura Angelescu McVey, Malgorzata Switek, Onnicha Sawangfa, and Jacqueline Smith Zweig. The happiness-income paradox revisited. *Proceedings of the National Academy of Sciences of the United States of America*, 107(52):22463–8, December 2010.

[15] Sara Beth Elson, Douglas Yeung, Parisa Roshan, S. R. Bohandy, and Alireza Nader. Using social media to gauge Iranian public opinion and mood after the 2009 election. Technical report, The RAND Corporation, 2012.

[16] Fairfax Media and Lateral Economics. The Herald/Age - Lateral Economics Index of Australia's Wellbeing. Technical Report December, Fairfax Media/Lateral Economics, 2012.

[17] Kevin R. Fontaine, Lawrence J. Cheskin, and Ivan Barofsky. Health-related quality of life in obese persons seeking treatment. *The Journal of Family Practice*, 43(3):265–270, 1996.

[18] Morgan R. Frank, Christopher M. Danforth, and Peter Sheridan Dodds. Mobility and happiness. *in preparation*, 2012.

[19] Gallup-Healthways. State of well-being 2011. Technical report, Gallup Inc., 2012.

[20] W. Hern. Urban malignancy: similarity in the fractal dimensions of urban morphology and malignant neoplasms. *International Journal of Anthropology*, 23:1–19, 2008.

[21] Daniel J. Hruschka. Do economic constraints on food choice make people fat? A critical review of two hypotheses for the poverty-obesity paradox. *American Journal of Human Biology*, 24(3):277–85, 2012.

[22] Jane Jacobs. *The Death and Life of Great American Cities*. Vintage Books, New York, 1961.

[23] Isabel M. Kloumann, Christopher M. Danforth, Kameron Decker Harris, Catherine A. Bliss, and Peter Sheridan Dodds. Positivity of the English language. *PloS one*, 7(1):e29484, January 2012.

[24] James A. Levine. Poverty and obesity in the U.S. *Diabetes*, 60(11):2667–8, November 2011.

[25] Andrew J. Oswald and Stephen Wu. Objective confirmation of subjective measures of human well-being: evidence from the U.S.A. *Science*, 327(5965):576–9, January 2010.

[26] L. A. Palinkas, D. L. Wingard, and E. Barrett-Connor. Depressive symptoms in overweight and obese older adults: a test of the "jolly fat" hypothesis. *Journal of*

Psychosomatic Research, 40(1):59–66, January 1996.

[27] Chris Randall. Measuring National Well-being - Where we Live - 2012. Technical Report July, UK Office for National Statistics, 2012.

[28] Jeffrey D. Sachs, Richard Layard, and John F. Hel-liwell. World Happiness Report. Technical report, Columbia University/ Canadian Institute for Advanced Research/London School of Economics, 2012.

[29] Guoqiang Shen. Fractal dimension and fractal growth of urbanized areas. *International Journal of Geographical Information Science*, 16(5):419–437, July 2002.

[30] Aaron Smith and Joanna Brenner. Twitter Use 2012. Technical report, Pew Research Institute, 2012.

[31] U.S. Census Bureau Demographic Internet Staff. Metropolitan and Micropolitan Statistical Areas. https://www.census.gov/programs-surveys/metro-micro.html , February 2013.

[32] Betsey Stevenson and Justin Wolfers. Economic Growth and Subjective Well-Being: Reassessing the Easterlin Paradox. *SSRN Electronic Journal*, 2008.

[33] Alois Stutzer. Limited self-control, obesity and the loss of happiness. *IZA Discussion Paper Report*, (2925), 2007.

[34] R. White and G. Engelen. Cellular automata and fractal urban form: a cellular modelling approach to the evolu-tion of urban land-use patterns. *Environment and Plan-ning A*, 25:1175–1199, 1993.

[35] Dan Witters. More than 15% obese in nearly all U.S. metro areas. http://www.gallup.com/poll/153143/obese-nearly-metro-areas.aspx, accessed February 2013.

[36] George Kingsley Zipf. *Human behavior and the principle of least effort*. Addison-Wesley, Reading, MA, 1949.

Appendices

Appendix A: Data set and states

In figure A1 we show the relationship between perimeter and area for the 3592 cities in the MAF/TIGER data database, which follow an approximate power law with exponent 1.294. The smallest city in both area and perimeter is Richmond, California, while the largest city is New York, whose perimeter extends far north into Connecticut and is agglomerated with Newark, New Jersey in this data set. We find that city area shows an approximate power-law dependence upon perimeter, with an average fractal dimension of $\alpha = 1.294$. Similar results have of course been reported previously for cities [29, 34], and has even been found to compare well with the fractal dimension of malignant skin lesions [20].

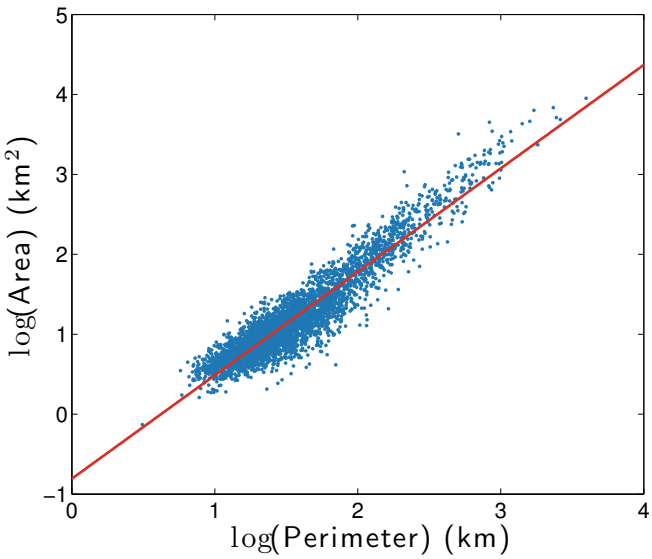

FIG. A1: Approximate power law relationship between city area and perimeter for all 3592 cities in the census data set. The fractal dimension is approximately 1.294.

In preprocessing the Twitter data set we have attempted to remove tweets from users that are clearly automated bots, in particular tweets from weather-recording services which periodically report values of temperature, humidity and the like. Users for whom more than 15% of their tweets contained the words 'humid', 'humidity', 'pressure' or 'earthquake' were removed from the dataset. We also made the decision to remove all variants of the racial pejorative or 'N-word' from calculations of h_{avg}. Variants of this word have very low happiness values, averaging $h_{\text{avg}} = 2.92$, and consequently were found to be highly influential in determining the average city happiness. However, when examining individual tweets we found that this word appeared to be being used in conversation as a more colloquial stand in for the word 'friend' in the vast majority of cases, and not in fact in any particularly negative sense. As such, we decided that scoring of the word was unfairly biasing our results towards the negative and removed it because of this. Future work will investigate the scoring of phrases instead of words, which will reduce the need for this type of adjustment.

For each city we create the normalized word frequency distribution $\hat{f}(i) = f_i/n$, where n is the total number of tweets collected for that city. The sum $\sum_i^N f_i/n$ therefore represents the average number of LabMT words per tweet, the mean of which is approximately 7.1. In figure A2 we show the average tweet length for the US cities for which we have collected more than 50000 words throughout 2011. Average tweet lengths range from 6.1 LabMT words per tweet for Orlando, Florida up to 7.8 words in Seattle, Washington.

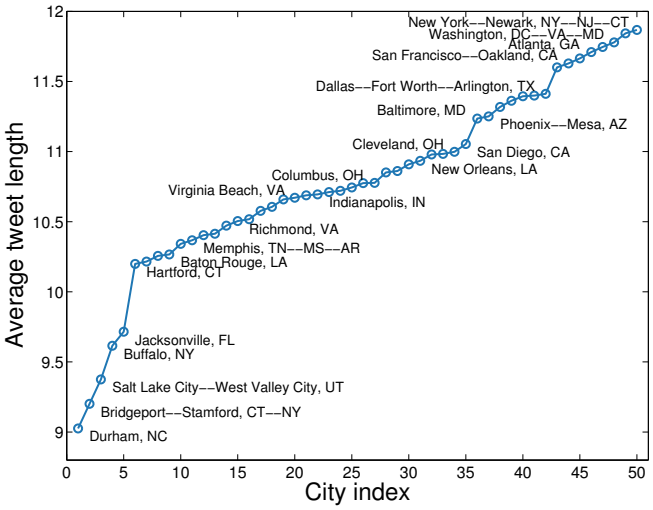

FIG. A2: Average message length for US cities with more than 50000 LabMT words collected during 2011.

Figure A3 shows choropleths for the number of geotagged tweets collected (left) and number of geotagged tweets normalized by state population (right) for the 2011 data set. In both plots the color scale is logarithmic. In table A1 we show the complete list of happiness scores for all US states. Word shift plots for each state are presented in Appendix B (online).

In tables A2 and A3 we show lists of the top 25 LabMT words with highest positive and negative correlation to obesity, respectively. In table A4 we show the words with lowest correlation to obesity, that is, the words with p-values greater than 0.9. Complete lists for for word correlations with all demographic attributes can be found in Appendix D (online) [1].

B,C,D,E,F Online appendices

The remaining appendices are located online, at http://www.uvm.edu/storylab/share/papers/mitchell2013a/. Appendix B contains word shift

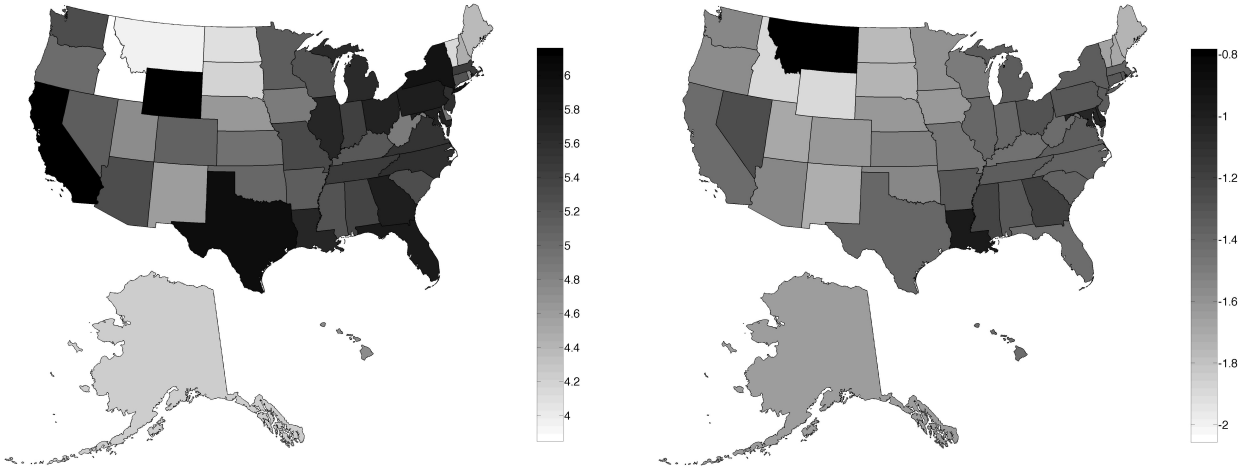

FIG. A3: Choropleths showing (the base-10 logarithm of) raw count (left) and number of geotagged tweets collected per capita (right) in each US state during the calendar year 2011.

plots for all states, Appendix C contains a comparison between happiness and the Gallup-Healthways well-being measure as well as tweet maps and word shifts for all cities, and Appendix D contains complete tables of correlations between demographic attributes and both happiness and word usage. Appendix E contains the complete list of LabMT words ordered by correlation with happiness, and Appendix F is a daily-updating happiness map of the United States.

Rank	State	h_{avg}
1	Hawaii	6.17
2	Maine	6.14
3	Nevada	6.12
4	Utah	6.11
5	Vermont	6.11
6	Colorado	6.10
7	Idaho	6.10
8	New Hampshire	6.09
9	Washington	6.08
10	Wyoming	6.08
11	Minnesota	6.07
12	Arizona	6.07
13	California	6.07
14	Florida	6.06
15	New York	6.06
16	New Mexico	6.05
17	Iowa	6.05
18	Oregon	6.05
19	North Dakota	6.04
20	Nebraska	6.04
21	Wisconsin	6.03
22	Kansas	6.03
23	Alaska	6.02
24	Oklahoma	6.02
25	Massachusetts	6.02
26	Montana	6.01
27	Missouri	6.01
28	Kentucky	6.00
29	New Jersey	5.99
30	West Virginia	5.99
31	Illinois	5.99
32	Rhode Island	5.99
33	Indiana	5.98
34	Texas	5.98
35	South Dakota	5.98
36	Virginia	5.97
37	Tennessee	5.97
38	Connecticut	5.97
39	Pennsylvania	5.97
40	South Carolina	5.96
41	North Carolina	5.96
42	Ohio	5.96
43	Arkansas	5.95
44	District of Columbia	5.94
45	Michigan	5.94
46	Alabama	5.94
47	Georgia	5.94
48	Delaware	5.92
49	Maryland	5.90
50	Mississippi	5.89
51	Louisiana	5.88

TABLE A1: Happiness scores h_{avg} for each US state, in order from highest to lowest.

Word	ρ	p-value	$h_{\text{avg}}(w_i)$
don't	0.461	2.28×10^{-11}	3.70
give	0.443	1.57×10^{-10}	6.54
lie	0.442	1.68×10^{-10}	2.60
hell	0.438	2.56×10^{-10}	2.22
my	0.438	2.74×10^{-10}	6.16
she	0.433	4.36×10^{-10}	6.18
okay	0.423	1.18×10^{-9}	6.56
like	0.419	1.72×10^{-9}	7.22
girl	0.419	1.76×10^{-9}	7.00
know	0.415	2.54×10^{-9}	6.10
act	0.412	3.48×10^{-9}	6.00
bitch	0.411	4.01×10^{-9}	3.14
me	0.403	8.5×10^{-9}	6.58
all	0.400	1.08×10^{-8}	6.22
nothin	0.399	1.14×10^{-8}	3.64
better	0.398	1.34×10^{-8}	7.00
bored	0.396	1.5×10^{-8}	3.04
bed	0.395	1.72×10^{-8}	7.18
sleep	0.395	1.78×10^{-8}	7.16
wish	0.388	3.25×10^{-8}	6.92
never	0.387	3.43×10^{-8}	3.34
money	0.380	6.41×10^{-8}	7.30
hate	0.378	7.57×10^{-8}	2.34
make	0.376	9.32×10^{-8}	6.00
cant	0.376	9.33×10^{-8}	3.48

TABLE A2: Top 25 words with strongest positive Spearman correlation ρ to obesity in 2011. Stop words with $4 < h_{\text{avg}} < 6$ have been removed from the list.

Word	ρ	p-value	$h_{\text{avg}}(w_i)$
cafe	-0.509	6.07×10^{-14}	6.78
photo	-0.493	4.87×10^{-13}	6.88
thai	-0.476	3.69×10^{-12}	6.22
fitness	-0.472	5.92×10^{-12}	6.92
park	-0.468	9.59×10^{-12}	7.08
yoga	-0.448	8.82×10^{-11}	7.04
restaurant	-0.448	8.93×10^{-11}	7.06
banana	-0.434	3.77×10^{-10}	6.86
event	-0.433	4.54×10^{-10}	6.12
hotel	-0.429	6.41×10^{-10}	6.16
spa	-0.420	1.54×10^{-9}	6.92
interesting	-0.420	1.62×10^{-9}	7.52
design	-0.409	4.76×10^{-9}	6.32
apple	-0.408	5.22×10^{-9}	7.44
feliz	-0.406	6.47×10^{-9}	6.04
photos	-0.404	7.8×10^{-9}	6.94
wine	-0.400	1.08×10^{-8}	6.42
bike	-0.399	1.22×10^{-8}	6.72
sun	-0.398	1.31×10^{-8}	7.80
delicious	-0.392	2.17×10^{-8}	7.92
flight	-0.391	2.34×10^{-8}	6.06
sunset	-0.391	2.51×10^{-8}	7.16
lounge	-0.389	2.93×10^{-8}	6.50
mortgage	-0.386	3.83×10^{-8}	3.88
dinner	-0.386	3.85×10^{-8}	7.40

TABLE A3: Top 25 words with strongest negative Spearman correlation ρ to obesity in 2011. Stop words with $4 < h_{\text{avg}} < 6$ have been removed from the list.

Word	ρ	p-value	$h_{\text{avg}}(w_i)$
olive	-0.001	9.94×10^{-1}	6.00
refrigerator	0.001	9.9×10^{-1}	N/A
hashbrowns	0.002	9.83×10^{-1}	N/A
eatting	-0.002	9.76×10^{-1}	N/A
sauteed	0.003	9.72×10^{-1}	N/A
fritos	-0.003	9.69×10^{-1}	N/A
munch	0.003	9.64×10^{-1}	N/A
doughnuts	-0.003	9.62×10^{-1}	N/A
cola	-0.004	9.62×10^{-1}	N/A
okra	-0.004	9.59×10^{-1}	N/A
grapes	0.004	9.51×10^{-1}	N/A
noodles	-0.004	9.51×10^{-1}	N/A
quiznos	0.005	9.49×10^{-1}	N/A
cucumbers	0.005	9.46×10^{-1}	N/A
chow	0.006	9.3×10^{-1}	N/A
walnut	0.007	9.28×10^{-1}	N/A
mulberry	0.007	9.19×10^{-1}	N/A
muesli	0.008	9.17×10^{-1}	N/A
hershey's	0.008	9.17×10^{-1}	N/A
snickers	0.008	9.16×10^{-1}	N/A
krispy	-0.008	9.15×10^{-1}	N/A
nugget	-0.008	9.12×10^{-1}	N/A
smores	0.008	9.1×10^{-1}	N/A
popcorn	0.009	9.07×10^{-1}	6.76

TABLE A4: The 24 food-related words which show least correlation with obesity, and have p-values greater than 0.9. Words are arranged in decreasing order of p-value.