9-12-2013

Happiness and the patterns of life: A study of geolocated tweets

Morgan R. Frank
University of Vermont

Lewis Mitchell
University of Vermont

Peter Sheridan Dodds
University of Vermont

Christopher M. Danforth
University of Vermont

Recommended Citation

Frank MR, Mitchell L, Dodds PS, Danforth CM. Happiness and the patterns of life: A study of geolocated tweets. Scientific reports. 2013 Sep 12;3(1):1-9.

SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
APPLIED PHYSICS
COMPLEX NETWORKS
FLUID DYNAMICS
QUALITY OF LIFE

Received
5 April 2013

Accepted
19 August 2013

Published
12 September 2013

Correspondence and requests for materials should be addressed to P.S.D. (peter.dodds@uvm.edu) or C.M.D. (chris.danforth@uvm.edu)

Happiness and the Patterns of Life: A Study of Geolocated Tweets

Morgan R. Frank, Lewis Mitchell, Peter Sheridan Dodds & Christopher M. Danforth

Computational Story Lab, Department of Mathematics and Statistics, Vermont Complex Systems Center, Vermont Advanced Computing Core, University of Vermont, Burlington, Vermont, United States of America.

The patterns of life exhibited by large populations have been described and modeled both as a basic science exercise and for a range of applied goals such as reducing automotive congestion, improving disaster response, and even predicting the location of individuals. However, these studies have had limited access to conversation content, rendering changes in expression as a function of movement invisible. In addition, they typically use the communication between a mobile phone and its nearest antenna tower to infer position, limiting the spatial resolution of the data to the geographical region serviced by each cellphone tower. We use a collection of 37 million geolocated tweets to characterize the movement patterns of 180,000 individuals, taking advantage of several orders of magnitude of increased spatial accuracy relative to previous work. Employing the recently developed sentiment analysis instrument known as the *hedonometer*, we characterize changes in word usage as a function of movement, and find that expressed happiness increases logarithmically with distance from an individual's average location.

A proper characterization of human mobility patterns[1–16] is an essential component in the development of models of urban planning[17], traffic forecasting[18], and the spread of diseases[19–21]. In the modern communication era, patterns of human movement have been revealed at an increasingly higher resolution in both space and time, with mobile phone data in particular complementing existing survey-based investigations. As is the case with each new instrument measuring macroscale sociotechnical phenomena, the task has become one of understanding what discernible patterns exist, and what meaning can be derived from those patterns[2,22–24].

Scientists working to understand mobility have employed a diverse set of methodologies. Brockmann et al.[7] used the circulation of nearly 1/2 million U.S. dollar bills whose locations were submitted by over 1 million visitors to a website[25] to demonstrate that bank note trajectories are superdiffusive in space and subdiffusive in time, i.e. moving farther and less frequently than expected.

Gonzalez et al.[1] used 6 months of mobile phone data from 100,000 individuals to show that human trajectories are regular in space and time, with each individual having a high probability of returning to a few preferred locations according to Zipf's law. Combining phone communication data with measures of community economic prosperity, Eagle et al.[2] showed that the diversity of contacts in an individual's social network is strongly correlated to the potential for economic development exhibited by their community. Finally, de Montjoye et al.[3] recently used mobile phone data to show that four space-time locations are enough to uniquely identify 95% of individuals.

Exemplifying recent work to characterize sentiment with social network communications, Mitchell et al.[26] combined traditional survey data (e.g., Gallup) with millions of tweets to correlate word usage with the demographic characteristics of U.S. urban areas. Expressed happiness was shown, for example, to correlate strongly with percentage of the population married, and anti-correlate with obesity. Words such as "McDonald's" and "hungry" appeared far more frequently in obese cities, suggesting their instrument could be used to provide real-time feedback on social health programs such as the proposed ban on the sale of large sodas in New York City in 2013.

In what follows, we characterize the pattern of life of over 180,000 individuals mainly in the U.S. using messages sent via the social networking service Twitter, and employ our text-based *hedonometer*[27] to characterize sentiment as a function of movement. In the calendar year 2011, we collected roughly 4 billion messages through Twitter's gardenhose feed, representing a random 10% of all status updates posted during this period.

Along with an abundance of other metadata, location information typically accompanies each message, resulting from one of three mechanisms by which individuals can report their location when updating their status. First, when an individual registers their account with Twitter, they are presented with the opportunity to report their location in a free text box. This region will be displayed in their user profile (e.g. 'NYC' or 'over the

rainbow'). The metadata accompanying each tweet sent by the individual contains this self-reported location. Second, individuals submitting a message through a web browser can choose to tag their message with a 'place' chosen from a drop-down menu, where the first option provided is typically the city within which the computer's IP address is found. For the purposes of accuracy, we have chosen to ignore each of these two mechanisms for reporting position when attempting to assign each tweet a geographical location, and focus instead on messages located via a third mechanism, namely the Global Positioning System (GPS).

Individuals using a mobile device application may opt-in to *geolocate* their message, in which case the exact latitude and longitude of the mobile phone is reported. The accuracy of this information is governed by the precision of the GPS instrument embedded in the phone, which can vary depending on the surrounding topography. As a result of these factors, we are able to approximately place each geolocated message inside a 10 meter circle on the surface of the Earth, within which the tweet was sent. Roughly 1% of the status updates received through the gardenhose feed are geolocated, resulting in a total of 37 million messages, collectively representing more than 180,000 English-speaking people worldwide. Fig. 1 illustrates the geospatial resolution of the data.

Results

Following González et al.[1], we examine the shape of human mobility using *radius of gyration*, hereafter gyradius, as a measure of the linear size occupied by an individual's trajectory. In Fig. 2, we investigate the geographical distribution of movement in four urban areas by plotting a dot for each tweet, colored by the gyradius of its author. Clockwise from the top left, cities are displayed in order of their apparent aggregate gyradius, with New York City seemingly exhibiting a smaller radius than the San Francisco Bay Area. In Chicago, many individuals writing from downtown exhibit an order of magnitude greater radius than individuals posting in areas outside of the city. A similar pattern is seen when looking at each point colored instead by distance from expected location (Fig. S2).

In the greater Los Angeles area, we see several clusters of individuals with larger radius in downtown Los Angeles, as well as Long Beach, Santa Monica, and Disneyland in Anaheim, while less densely populated areas are seen as smaller clusters exhibiting much smaller radii. The geography of the San Francisco Bay Area is clearly revealed, with many large radius individuals tweeting from downtown San Francisco, and somewhat less homogeneity in Oakland and San Jose. Outside of these cities, there are many suburban areas revealed by individuals with large radius, e.g. Palo Alto. Tweets appearing in less densely populated Bay Area locations appear to be far more likely to be authored by large radius individuals than those appearing in lower population areas of the other cities. This observation likely reflects the socio-economic and demographic characteristics of individuals using Twitter in the Bay Area, where the social network service was founded. Additionally, it could reflect the presence of tourists who will typically have a larger radius than someone who lives and works in the Bay Area.

We calculate Geary's C (local) and Moran's I (global) spatial autocorrelation for the data shown in Figures 2 and S2, finding statistical support for spatial clustering in each (Tables S1 and S2). However, the correlations benefit from the propensity for each individual's collection of tweets to exhibit clustering. To avoid this confound, we also make city plots of mode location colored by gyradius, where each dot represents an individual rather than a tweet. These figures are not included to respect the privacy of individuals in the study. Table S3 reports the strong spatial autocorrelation we observed, reflecting a form of geospatial homophily: the tendency of individuals to author messages in proximity to others with similar gyradius. Tourists are unlikely to be included in this statistic, given the nature of mode location, and as such the clustering is potentially a result of similar commute distances.

One observation seemingly apparent in Fig. 2 is that individuals who move a lot tend to appear in areas of large population density. Given the apparent economies of scale offered by living in a densely populated area, one might expect to observe the inverse relationship, namely that people living in less densely populated areas travel further, by necessity, to their place of employment or grocery store, for

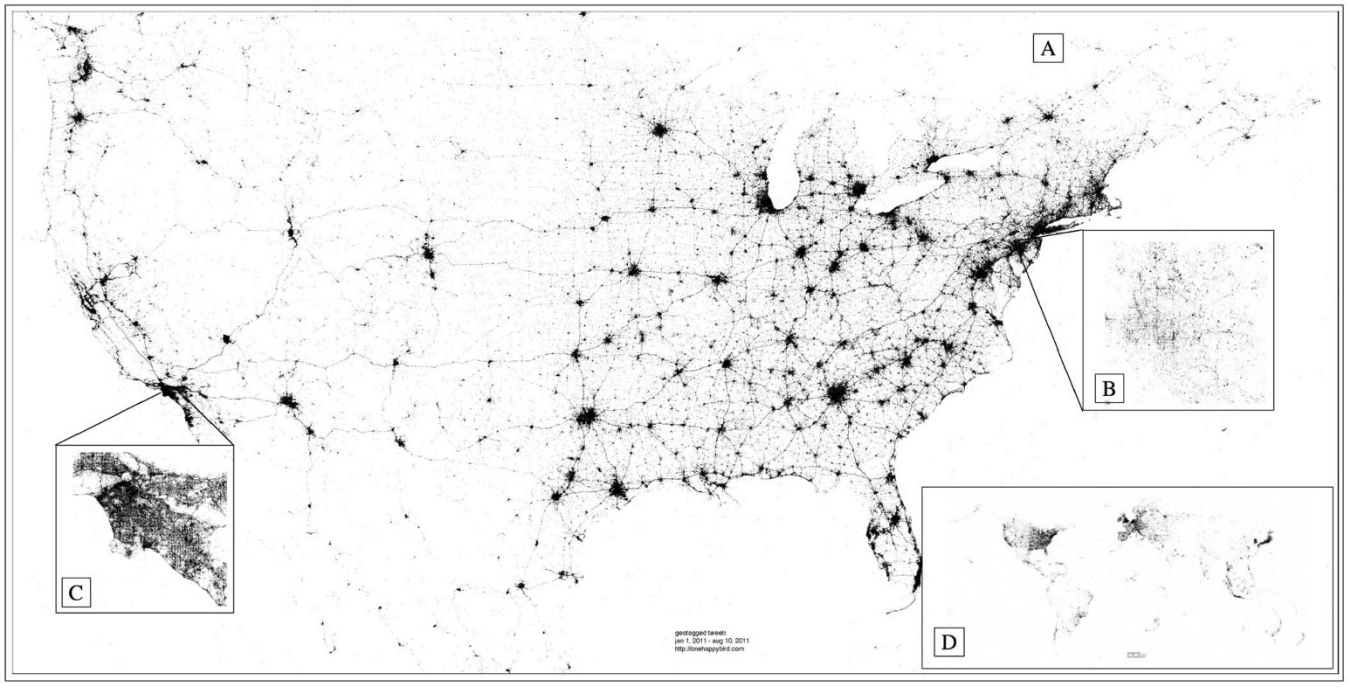

Figure 1 | Each point corresponds to a geolocated tweet posted in 2011. Twitter activity is most apparent in urban areas. Note that the image contains no cartographic borders, simply a small dot for each message. Legend: A (U.S.), B (Washington, D.C.), C (Los Angeles, C.A.), and D (Earth). Maps were created using Matlab.

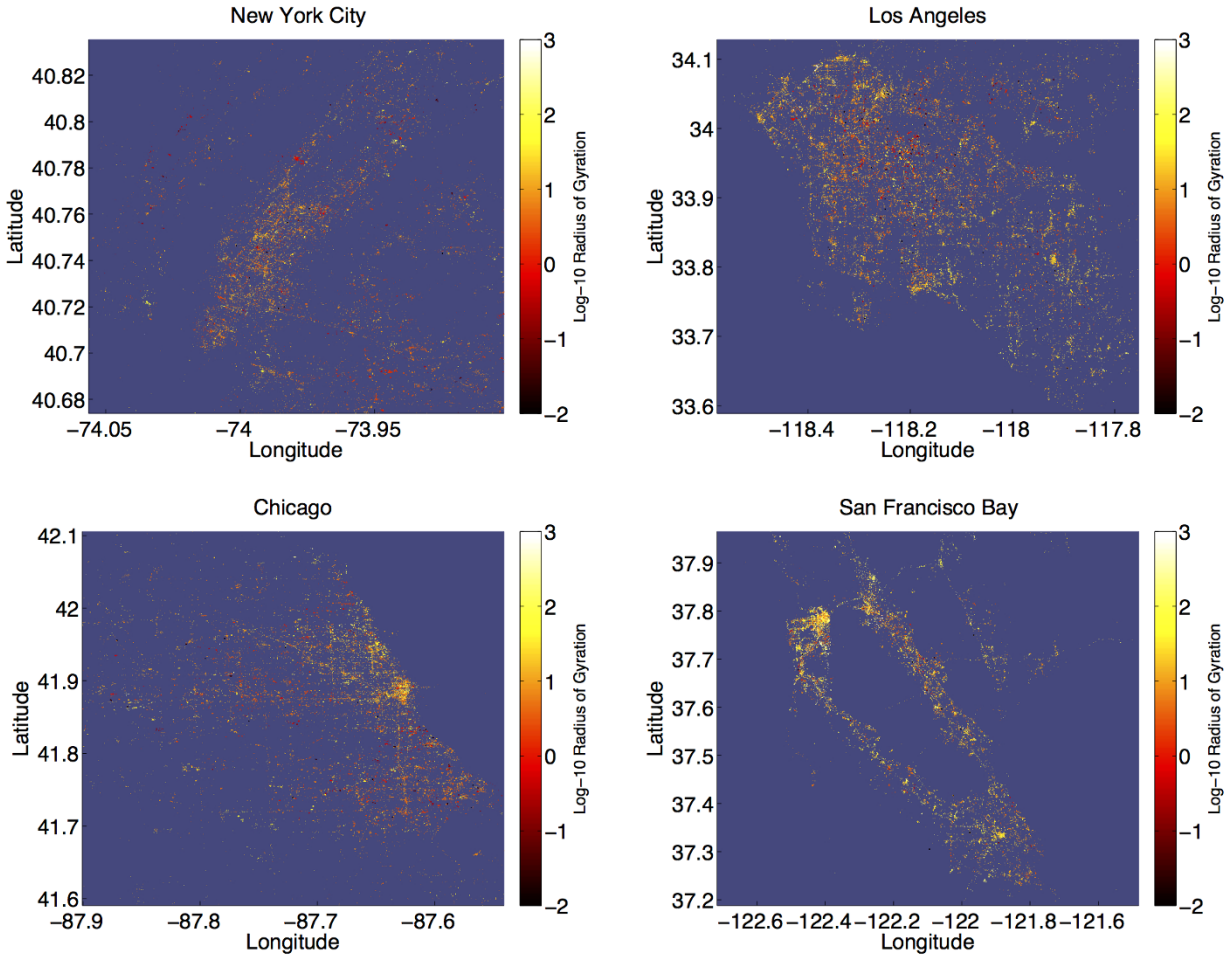

Figure 2 | The gyradius, calculated for each individual, is shown for each tweet authored in four example cities. Tweet activity reflects population density, with urban areas clearly visible in each city. Histograms of gyradii for each city are shown in Fig. S1, along with tweet locations colored by distance from expected location (Fig. S2). The number of tweets shown for each city is $N = 56650$ (Chicago), $N = 103{,}213$ (Los Angeles), $N = 42{,}089$ (New York City), and $N = 45{,}754$ (San Francisco). Note that higher resolution versions of the four panels above can be found online[28]. Maps were created using Matlab.

example. Of course, individuals observed to have a large radius could be tourists, or they could have a long commute. Nevertheless, we find no statistical evidence for this trend. Comparing individuals whose average location falls in an area of small vs. large tweet density, we observe little difference in their average gyradii (not shown).

Moving beyond these four urban areas and looking at 472 cities in the U.S., we do find a moderate correlation between the mean gyradius and city land area (Pearson $\rho = 0.24$, $p = 2 \times 10^{-7}$); Fig. S3 and Table S4 show the top and bottom cities with respect to gyradii.

To investigate the shape of human mobility, we normalize each individual's trajectory to a common reference frame (see Methods). In Fig. 3, we plot a heat map of the probability density function of the normalized locations of all individuals. For the purposes of this discussion, we will refer to deviations from an individual's expected location in the normalized reference frame as occurring in the directions north, south, east, and west. Several features of the map reveal interesting patterns of movement. First, the overall west-to-east teardrop shape of the contours demonstrates that people travel predominantly along their principle axis, namely heading west from the origin along $y/\sigma_y = 0$, with deviations in the orthogonal direction becoming shorter and less frequent as they move farther away from the origin.

Second, the appearance of two spatially distinct yellow regions separated by a less populated green region suggests that people spend the vast majority of their time near two locations. We refer to these locations as the *work* and *home* locales[8], where the home locale is centered on the dark red region roughly 1 standard deviation east of the origin, and the work locale is centered approximately 2 standard deviations west of the origin. These locations highlight the bimodal distribution of principal axis corridor messages (Fig. 4A).

Finally, a clear asymmetry is observed about the $x/\sigma_x = 0$ axis indicating the increasingly isotropic variation in movement surrounding the home locale, as compared to the work locale. We interpret this to be a reflection of the tendency to be more familiar with the surroundings of one's home, and to explore these surroundings in a more social context (Fig. 4B). The symmetry observed when reflecting about the $y/\sigma_y = 0$-axis is strong, demonstrating the remarkable consistency of the movement patterns revealed by the data.

In an effort to characterize the temporal and spatial structure observed in Fig. 3, in Fig. 5 we examine locations frequently visited by the most active members of our data set, namely the roughly 300 individuals for whom we received at least 800 geolocated messages. We suspect that these individuals enabled the geolocating feature to be *on* by default for all messages, as implied by the roughly $O(10^4)$ geolocated messages suggested by the gardenhose rate. In Fig. 5, we focus on these individuals specifically; of all participants, their prolific tweet activity most accurately reflects their movement profile.

The main figure shows the probability of tweeting from each locale, with locales ordered by rank, for each individual[8]. We find that $P\left(H_i^{(a)}\right) \propto R\left(H_i^{(a)}\right)^{-1.3}$ which is approximately a Zipf distribution[29]. This finding indicates that regardless of the number of tweet

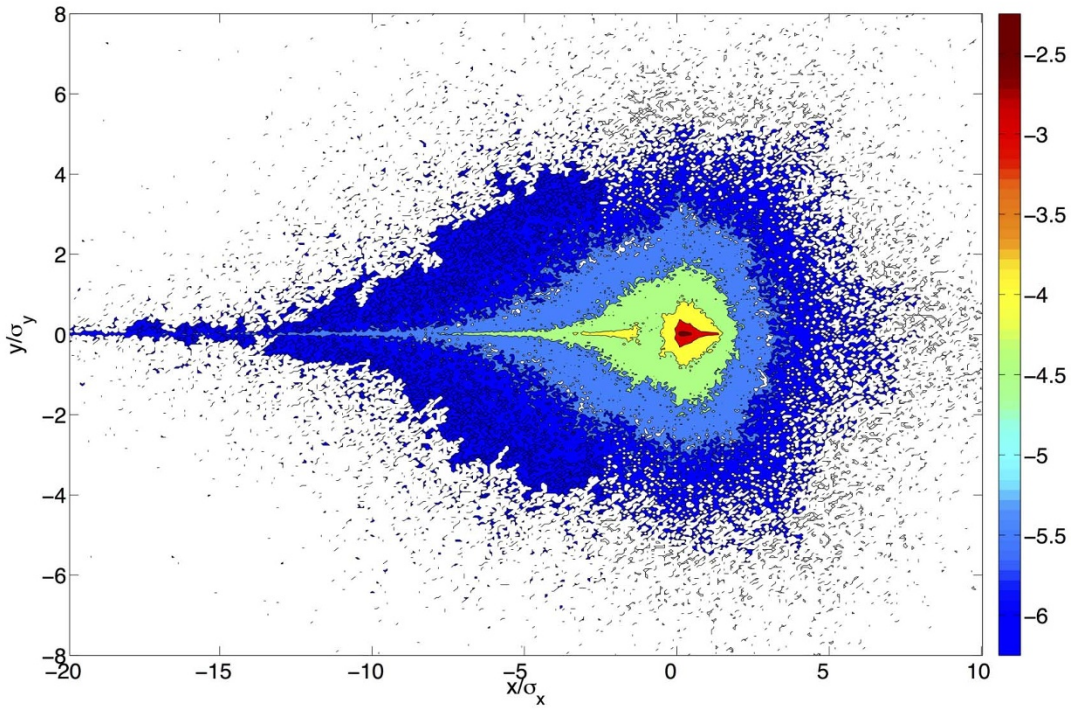

Figure 3 | The probability density function of observing an individual in their normalized reference frame, where the origin corresponds to each individual's expected location, and $\sigma_y = 0$ corresponds to their principle axis. This map shows the positions of over 37,000 individuals, each with more than 50 locations, in their intrinsic reference frame.

locales for a given individual, the majority of their messaging activity occurs in one of only a few locales, with the probability decaying at a predictable rate. If the decay were Zipfian, an individual would be approximately *n*-times as likely to tweet from their mode location than from their rank *n* location. With our slope being steeper, these probabilities fall at a faster rate with rank. The slope is robust to variation in number and composition of individuals (Fig. S4).

For roughly 95% of these individuals, each tweet has a greater than 10% chance of being authored from their mode location (Fig. 5B). Fig. 5C demonstrates each individual's likelihood of authoring messages from their mode location (black curve) at different times of day throughout the week. A period-2 cycle is observed for each day of the week. Maxima are seen in the morning (8–10am) and evening (10pm-midnight), and minima in the afternoon (2–4pm) and overnight (2–4am) hours. The peak in the morning is consistently higher than that in the evening, and the afternoon valley is consistently lower than the overnight valley. The cycle is somewhat less structured on the weekend. Also plotted are the probabilities of tweeting from locations other than the mode (red curve).

In a study performed with cellphone tower data, González et al.[1] found that people spend most of their time in two locations, and a person's probability of being found at a separate location diminishes

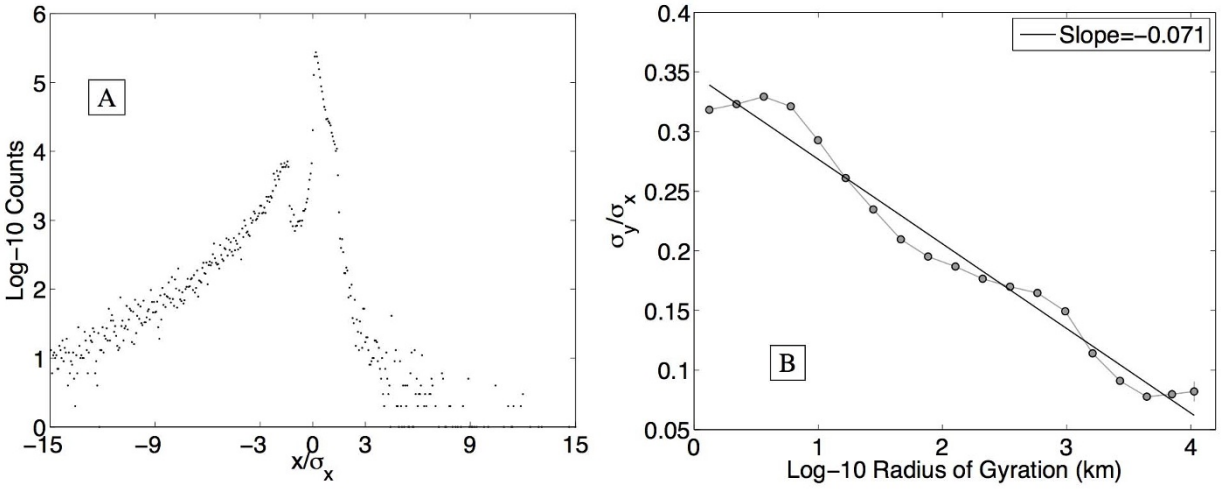

Figure 4 | Looking at messages authored in the principle axis corridor, defined by $\left|\frac{y}{\sigma_y}\right| < \frac{30}{1000}$, we observe a clear separation between the most likely and second most likely position (A). The distribution is skewed left, with movement in a heading opposite an individual's work/home corridor observed to be highly unlikely. In addition, due to the normalization, we see that individuals are much more likely to tweet slightly east of their expected location than slightly west. The isotropy ratio (B) measures the change in the density's shape as a function of gyradius, with large radius individuals exhibiting a less circular pattern of life. Standard errors are plotted, but are only visible for the largest radius group. The isotropy ratio decays logarithmically with radius.

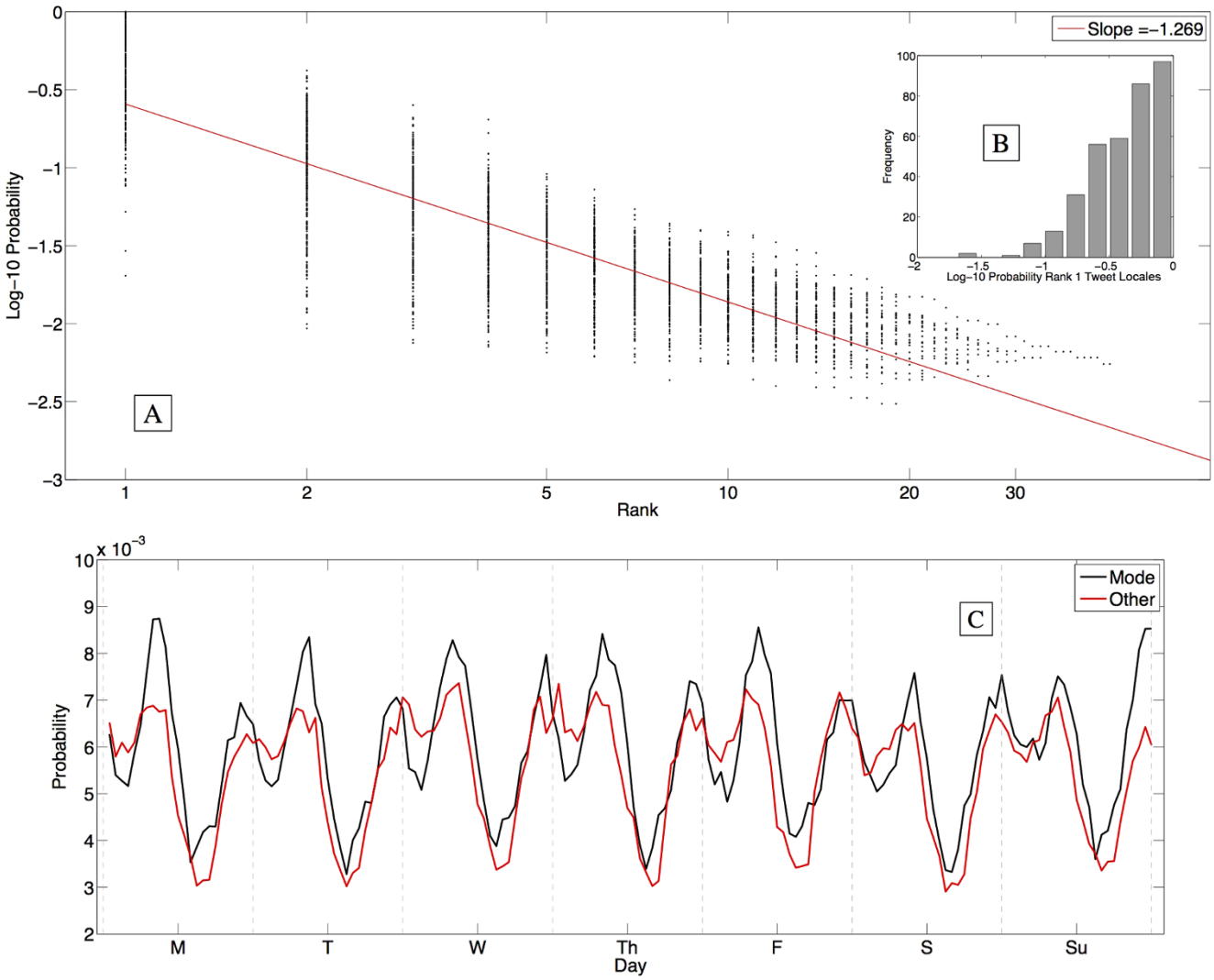

Figure 5 | Representing the approximately 300 individuals for whom we have at least 800 geolocated messages, we plot the probability of tweeting from a locale as a function of the tweet locale rank (A). Each dot represents a single individual's likelihood of tweeting from one of their locales. The axes are logarithmic, revealing an approximate Zipfian distribution with slope −1.3[29]. (B) Distribution of the rank-1 locale, each individual's mode location. (C) A robust diurnal cycle is observed in the hourly time of day at which statuses are updated, with those from the mode location (black curve) occurring more often than other locations (red curve) in the morning and evening. Probabilities sum to 1 for each curve, with bins for each hour. Dashed vertical lines denote midnight.

rapidly with rank by visitation. While our investigation reveals a similar pattern, we find a larger difference in the probability that an individual is tweeting from the home locale than from the work locale. We attribute these slight differences in our results to the different spatiotemporal precision of location data, as well as differences in activities represented by the data. González et al. determined each individual's location by continuously monitoring the nearest cellphone tower whose range they were within. As such, we receive more precise location information, but only when individuals performed the act of tweeting.

One major advantage of using Twitter data to study movement is the additional source of information provided by the messages themselves. Researchers using mobile phone data to characterize mobility patterns do not have access to conversations occurring during the time period of interest. To measure the sentiment associated with different patterns of movement, we use the hedonometer introduced by Dodds et al.[27]. The instrument performs a context-free measurement of the happiness of a large collection of words using the language assessment by Mechanical Turk (labMT) word list, as described in Kloumann et al.[30]. LabMT comprises roughly 10,000 of the most frequently used words in the English language, each of which was scored for happiness on a scale of 1 (sad) to 9 (happy) by people using Amazon's Mechanical Turk service[31,32], resulting in an average happiness score for each word. Example word scores are shown in Table 1. Note that in employing the hedonometer, we avoid assigning sentiment to individual tweets, a challenging task more appropriately suited to advanced natural language processing software.

To examine the relationship between movement and happiness, we calculate expressed happiness as a function of distance from an individual's expected location, as well as gyradius. For the former, we grouped tweets into ten equally populated bins, with each group containing more than 500,000 tweets from similar distances. The happiness of each group was then computed using Eqn 3 (see Methods), where all words written from a given distance were gathered into a single bin. For the latter, we placed individuals into ten equally sized groups by gyradius, with each group containing more than 10,000 individuals with similar gyradii.

Fig. 6 plots average word happiness against the distance from expected location (A), and gyradius (B). Starting with location, we

Table 1 | Example language assessment by Mechanical Turk (labMT)[27,30] words and scores. Words with neutral scores $4 < h_{avg}(w_i) < 6$ are colored gray and ignored when assigning the happiness score to a large text

word	$h_{avg}(w_i)$
'happy'	8.30
'hahaha'	7.94
'fresh'	7.26
'cherry'	7.04
'pancake'	6.96
'piano'	6.94
'and'	5.22
'the'	4.98
'of'	4.94
'down'	3.66
'worse'	2.70
'crash'	2.60
':('	2.36
'war'	1.80
'jail'	1.76

find that tweets written close to an individual's center of mass are slightly happier than those written 1 km away. The least happy words, on average, are used at a distance representative of a short daily commute to work. Beyond this least happy distance, remarkably we find that happiness increases logarithmically with distance from expected location. Perhaps even more remarkably, we find an almost identical trend when grouping together individuals rather than tweets, observing that happiness also increases logarithmically with gyradius. Individuals with a large radius use happier words than those with a smaller pattern of life. We find the trend observed in Fig. 6 holds for 3 of the 4 urban areas (Los Angeles, San Francisco, and Chicago), see Figs. S5, S6.

To explain the difference in expressed happiness exhibited by different mobility groups, we turn to word shift graphs in Fig. 7. Word shift graphs were introduced by Dodds and Danforth[27,33] as a means for investigating the elements of language responsible for happiness differences between two large texts. As an example, consider the difference between tweets authored at distances of roughly 1 km and 2500 km away from an individual's expected location. The average happiness scores for these two distances are $h_{avg} = 5.96$ and $h_{avg} = 6.13$ respectively. Individual word contributions to this difference are shown in Fig. 7A, and can be described as follows.

Words appearing on the right increase the happiness of the 2500 km distance relative 1 km distance. For example, tweets authored far from an individual's expected location are more likely to contain the positive words 'beach', 'new', 'great', 'park', 'restaurant', 'dinner', 'resort', 'coffee', 'lunch', 'cafe', and 'food', and less likely to contain the negative words 'no', 'don't', 'not', 'hate', 'can't', 'damn', and 'never' than tweets posted close to home. Words going against the trend appear on the left, decreasing the happiness of the 2500 km distance group relative to the 1 km group. Tweets close to home are more likely to contain the positive words 'me', 'lol', 'love', 'like', 'haha', 'my', 'you', and 'good'. Moving clockwise, the three insets in Fig. 7A show that the two text sizes are comparable, the biggest contributor to the happiness difference is the decrease in negative words authored by individuals very far from their expected location, and the 50 words listed make up roughly 50% of the total difference between the two bags of words.

Note that the relatively small differences in h_{avg} scores reflect a small signal, yet one that we have shown previously can be resolved by our hedonometer[27]. Additional word shift comparisons for the four urban areas investigated earlier are provided in the Supplemental Material, Figs. S7, S8.

Looking at the word differences between individuals with largest and smallest radii of gyration in Fig. 7B, we see that individuals in the large radius group author the negative words 'hate', 'damn', 'dont', 'mad', 'never', 'not' and assorted profanity less frequently, and the positive words 'great', 'new', 'dinner', 'hahaha', and 'lunch' more frequently than the small radius group. Going against the trend, the large radius group uses the positive words 'me', 'lol', 'love', 'like', 'funny', 'girl', and 'my' less frequently, and the negative words 'no', and 'last' more frequently. Comparing with other groups, the large radius group authors an increased frequency of words in reference to eating, like the words 'dinner', 'lunch', 'restaurant', and 'food', and make less reference to traffic congestion.

Comparing the two figures, we note that individuals with large radius laugh more (e.g 'hahaha') than those with a small radius, but individuals closer to their expected location laugh more than those far from home.

These word differences reveal the relationship between an individual's pattern of movement and their experiences. It is not surprising

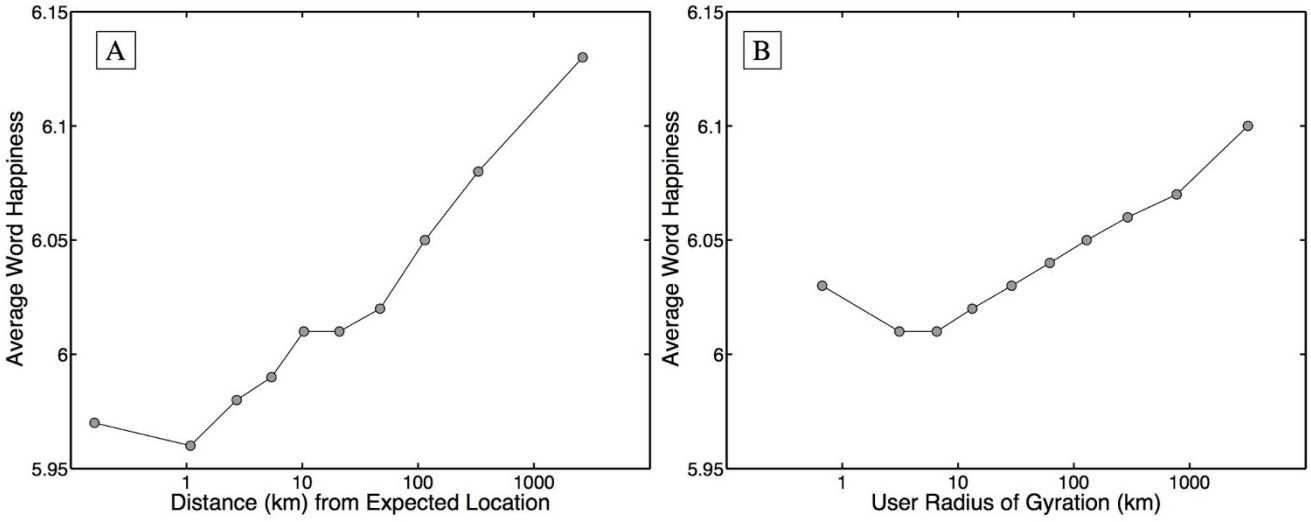

Figure 6 | (A) Average happiness of words written as a function of distance from an author's expected location, with tweets grouped into ten equally populated bins. Expressed happiness grows logarithmically with distance distance from expected location. (B) A similar trend is observed when individuals are grouped into ten equally populated bins according to their gyradius. Both trends persist through variations in binning and different measures of mobility.

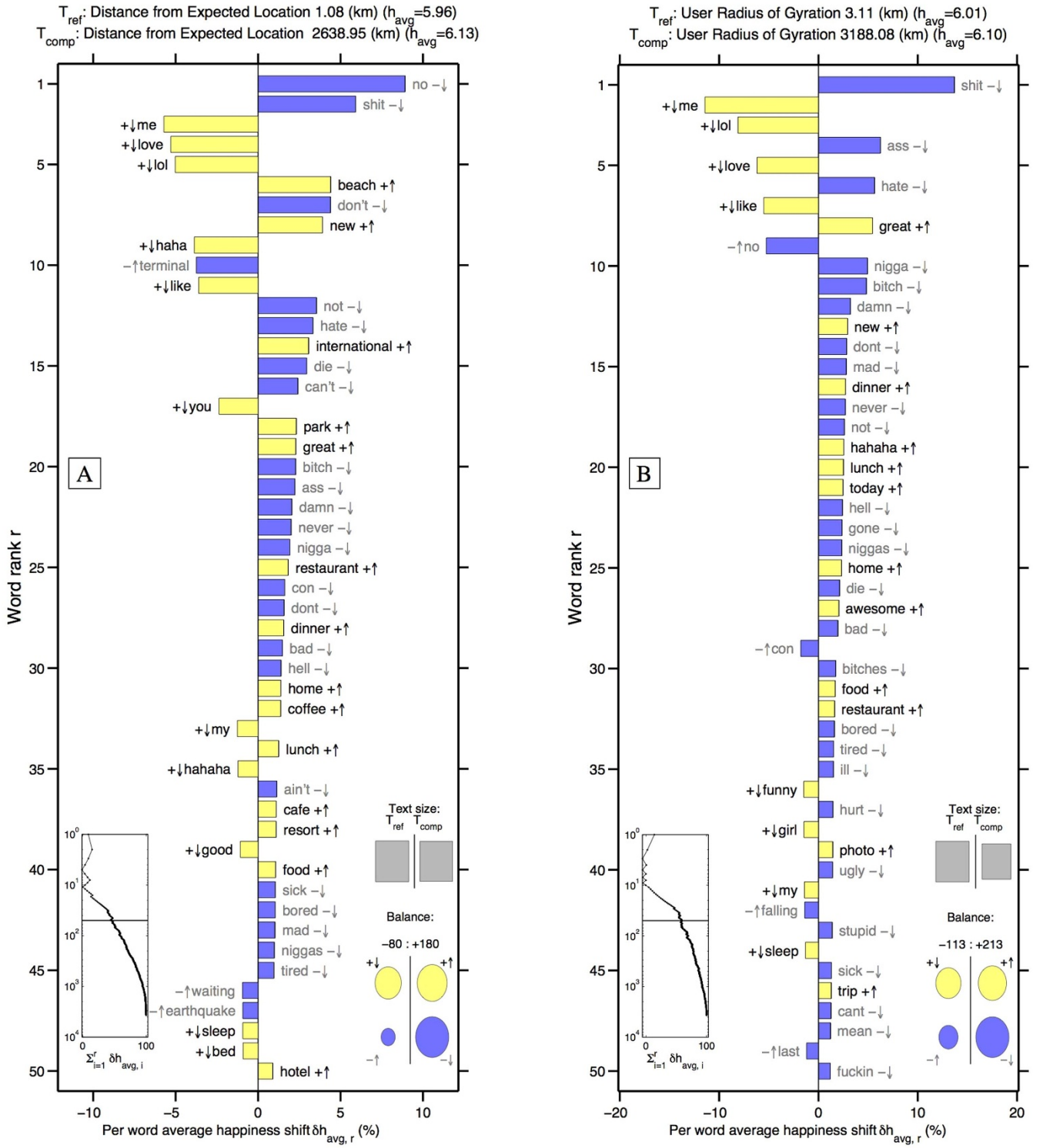

Figure 7 | Word shift graphs comparing (A) the lowest average word happiness distance from home group to the words authored farthest from home, which also has the largest average word happiness and (B) the smallest gyradius group with the largest gyradius group. The words in the word shifts from top to bottom appear in decreasing order of ranked percentage contribution to the overall average happiness difference (Δh_{avg}) of the two texts being compared. The +/− symbols indicate whether the word has an average happiness score that is happy or sad relative to the entire text T_{ref}. The symbols ↑/↓ indicate whether a word was used more or less in T_{comp} relative to usage in T_{ref}. The left inset panel shows how the ranked top contributing words to Δh_{avg} combine in sum. The four circles in the lower right show the total contribution of the four word types (+ ↑, − ↑, + ↓, − ↓) to the Balance of the happiness difference. The number of words in each of the two texts is represented by the relative area of the grey squares (Text size). See Dodds et al.[27] for further details and examples of word shift graphs.

to observe regular international travelers tweeting about the food they enjoy on vacation. Indeed, we expect that individuals capable of tweeting at a great distance from their expected location are more likely to benefit from an advantaged socioeconomic status, which they happily update frequently. In our earlier work, we have demonstrated that expressed happiness correlates strongly with many socioeconomic indicators[26]. Nevertheless, setting aside these luxurious words, we still see a general decline in the use of negative words as individuals travel farther from their expected location. In fact, of the four contributions to the difference in happiness between words authored close to home vs. far from home, this decline in negative words is the largest component (bottom right inset, Fig. 7).

Discussion

Using 37 million geolocated tweets authored in 2011, we have been able to characterize the pattern of life of over 180,000 individuals largely residing in the United States. While observed mobility patterns agree qualitatively with previous work investigating cellphone data[1], we are able to connect movement patterns to changes in word usage for the first time. Our main finding is that expressed happiness increases logarithmically with both distance from expected location and gyradius, largely because individuals who travel farther use positive, food related words more frequently, and negative words and profanity less frequently.

Several methodological issues are raised by the use of Twitter messages to characterize mobility and happiness. Considering Twitter as a source, we note that according to the Pew Internet & American Life Project, roughly 15% of adults in the U.S. were actively using Twitter at the end of 2011[34]. While this fraction represents a substantial group of Americans, we have no data to quantify the demographic group represented by the subset of these 15% who specifically choose to geolocate a large percentage of their messages. Nevertheless, since we threshold the sample to include individuals who have geolocated more than approximately 300 of their messages in 2011, we suspect that the large majority of individuals represented in our study regularly do so as a matter of daily life, as opposed to geolocating messages only when encountering a novel experience such as a vacation.

Regarding word usage as a proxy for happiness, accessing the internal emotional state of individuals is beyond the scope of our instrument. We do believe however, that when aggregated, the words used by large groups of individuals reflect their culture in ways not captured by surveys or self-report. Indeed, we see the hedonometer as complementing more traditional economic methods for characterizing economic and societal health, such as the Gross Domestic Product or Consumer Confidence Index. Using the same collection of geolocated messages explored here, the hedonometer was recently employed by Mitchell et al.[26] to characterize trends in word usage for cities. Expressed happiness was shown to correlate to hundreds of demographic, socio-economic, and health measures, with interactive evidence available in the article's online Appendix[35].

Our work contributes to a growing body of literature aimed at observing, describing, modeling, and ultimately explaining the spatiotemporal dynamics of large-scale socio-technical systems. The mobility patterns investigated here could be combined with more traditional surveys (e.g. census data) to inform public policy regarding many important issues, for example relating to the 'obesity epidemic' and changes in word usage at the level of individual neighborhoods targeted by public health campaigns. Feedback on society's eating behavior in response to health promotion policies could be available at the level of neighborhoods on a time scale of weeks, in advance of health data outcomes that typically take years. Indeed, epidemiological models of the spread of food-borne illness can now concurrently leverage information about social network connections and geographic proximity[36].

In addition, future mental health providers could flag changes in individual behavior revealed through patterns of movement and communication for intervention. For example, a depressed emotional state may be indicated by simultaneously observing marked declines in gyradius, decreased social interactions, and sustained increase in usage of negative words. Natural extensions of this work might combine topological measures of network interactions with geospatial data to predict the likelihood of new links appearing in a social network[37], or to measure the spread of emotions through geographical and topological space[38].

Methods

In an effort at quality control for the geolocated messages, we identified and removed messages posted by robotic accounts and programmed tweeting services designed to automatically send tweets typically not reflecting information about human activity. Preliminary analyses revealed a noticeable presence of bots posting geolocated messages referring to weather, earthquakes, traffic, and coupons. We identified and ignored tweets collected from individuals for whom at least half of their tweets contained any of the words 'pressure', 'humid', 'humidity', 'earthquake', 'traffic' or 'coupon'.

Messages referencing Foursquare check-ins (typically of the form 'I'm at starbucks http://4sq.com/qrel9d') were retained for the purpose of characterizing the mobility profile of each individual. However, for results involving happiness, we ignored Foursquare check-in tweets as their content is unlikely to directly reflect sentiment.

Finally, to ensure that individual movement profiles are based on a reasonably sized collection of locations, for this study we focus on individuals for whom we have at least 30 geolocated tweets. Given the uniformity of the random sample provided by the gardenhose, we can assume these individuals geolocated a minimum of approximately 300 status updates in 2011. Individuals were included in Figures 6, 7 if their messages matched LabMT words.

For reasons of privacy, we ignored all user specific information including individual names. In addition, where the trajectories traced out by specific individuals are visualized, we obscured the coordinate system of reference. Tweets were assigned to urban areas as defined by the 2010 United States Census Bureaus MAF/TIGER (Master Address File/Topologically Integrated Geographic Encoding and Referencing) database[39].

The gyradius for individual a is defined as

$$r^{(a)} = \sqrt{\frac{1}{N^{(a)}} \sum_{i=1}^{N^{(a)}} \left(\vec{p}_i^{(a)} - \langle \vec{p}^{(a)} \rangle \right)^2} \qquad (1)$$

where the two-dimensional vector $\vec{p}_i^{(a)}$ is the ith position in the trajectory of individual a, given by the geolocation of that individual's ith tweet, as observed in our database. $N^{(a)}$ is the total number of tweets from individual a, and $\langle \vec{p}^{(a)} \rangle = 1 \Big/ N^{(a)} \sum_{i=1}^{N^{(a)}} \vec{p}_i^{(a)}$ is the center of mass of their trajectory, which we denote their *expected location*. Note that if we consider each message to be a prediction of an individual's location, then the gyradius is in fact the root mean square error (RMSE) of that prediction. Fig. S9 plots the Complementary Cumulative Distribution Function (CCDF) of the gyradii of all individuals.

To compare the shape of individual trajectories, we normalize for both differences in gyradius and direction of trajectory. Considering each individual's trajectory as a set of (x,y)-pairs $\{(x_1,y_1), (x_2,y_2),\ldots,(x_N,y_N)\}$, we calculate the two dimensional matrix known as the tensor of inertia, considering each point in a individual's trajectory as an equally weighted mass at location (x_i,y_i). We then find this tensor's eigenvectors and eigenvalues. The eigenvector corresponding to the largest eigenvalue represents the axis along which most of the individual's trajectory occurs (hereafter called the individual's *principal axis*). Previous work has demonstrated that for most individuals, this axis is parallel to the corridor between their work location and home[1,4].

To normalize the different compass orientations of individual trajectories, we rotate the coordinate system of each individual so that their principal axis points due west. The expected location for each individual $(\bar{x}, \bar{y})$ is then used to translate their position vector, i.e. $(x_i - \bar{x}, y_i - \bar{y})$, to ensure that the shape of each individual's trajectory is in a common frame of reference. However, the distances travelled by each individual vary widely despite their shared orientation (e.g. pedestrian vs. airline commute). In order to compare these trajectories, we calculate the standard deviation σ_x, σ_y for a given individual's trajectory, and divide their x- and y-coordinates by σ_x and σ_y, respectively. For more information about this process, including a pair of example trajectory normalizations, see Figs. S10–S14.

In an attempt to characterize time spent in each location, we define the ith tweet *locale* for individual a, denoted $H_i^{(a)}$, to be a circle within which individual a posted at least 10 messages[8]. The center of the circle is defined by the average position of all messages appearing in the locale, and the radius of the circle is chosen such that each tweet posted within a locale is at most 100 meters away from the center, and no locales overlap. To measure the importance of locale i to individual a, we count the number of messages appearing in each tweet locale and produce the ranking $R\left(H_i^{(a)}\right)$ for individual a. The probability that individual a tweets from locale $H_i^{(a)}$ is

$$P\left(H_i^{(a)}\right) = \frac{\left|H_i^{(a)}\right|}{N^{(a)}} \quad (2)$$

where $\left|H_i^{(a)}\right|$ is the number of tweet locations contained in $H_i^{(a)}$. Notice that the locale probabilities for individual a may not sum to one since it may be the case that individual a has tweet locations that are not contained in a tweet locale. Hereafter, we will refer to an individual's most frequently visited, or rank-1 locale, as their mode location.

Using the labMT scores[27], we determine the average happiness (h_{avg}) of a given text T containing N unique words by

$$h_{avg}(T) = \frac{\sum_{i=1}^{N} h_{avg}(w_i)\cdot f_i}{\sum_{i=1}^{N} f_i} = \sum_{i=1}^{N} h_{avg}(w_i)\cdot p_i \quad (3)$$

where f_i is the frequency with which the ith word w_i, for which we have an average word happiness score $h_{avg}(w_i)$, occurred in text T. The normalized frequency of w_i is then given by $p_i = f_i/\sum_{i=1}^{N} f_i$.

The hedonometer instrument can be tuned to emphasize the most emotionally charged words by removing words within Δh_{avg} of the neutral score of $h_{avg} = 5$. We have further shown that ignoring these neutral words with $4 < h_{avg}(w_i) < 6$ provides a good balance of sensitivity and robustness, and thus we chose $\Delta h_{avg} = 1$ for this study[27].

1. González, M. C., Hidalgo, C. A. & Barabási, A. L. Understanding individual human mobility patterns. *Nature* **453**, 779–782 (2008).
2. Eagle, N., Macy, M. & Claxton, R. Network Diversity and Economic Development. *Science* **328**, 1029–1031 (2010).
3. de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M. & Blondel, V. D. Unique in the Crowd: The privacy bounds of human mobility. *Sci. Rep.* **3** (2013).
4. Song, C., Qu, Z., Blumm, N. & Barabsi, A.-L. Limits of Predictability in Human Mobility. *Science* **327**, 1018–1021 (2010).
5. Simini, F., González, M. C., Maritan, A. & Barabási, A.-L. A universal model for mobility and migration patterns. *Nature* **484** (2012).
6. Wang, D., Pedreschi, D., Song, C., Gionatti, F. & Barabási, A.-L. Human mobility, social ties, and link prediction. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1100–1108 (2011).
7. Brockmann, D. D., Hufnagel, L. & Geisel, T. The scaling laws of human travel. *Nature* **439**, 462–465 (2006).
8. Bagrow, J. P. & Lin, Y-R. Mesoscopic Structure and Social Aspects of Human Mobility. *PLoS ONE* **7**(5), e37676 (2012).
9. Ramos-Fernandez, G. *et al*. Lévy walk patterns in the foraging movements of spider monkeys (Ateles geoffroyi). *Behav. Ecol. Sociobiol.* **55**, 223–230 (2004).
10. Palla, G., Barabási, A.-L. & Vicsek, T. Quantifying social group evolution. *Nature* **446**, 664–667 (2007).
11. Hidalgo, C. A. & Rodriguez-Sickert, C. The dynamics of a mobile phone network. *Physica A.* **387**, 3017–3024 (2008).
12. Barabási, A.-L. The origin of bursts and heavy tails in human dynamics. *Nature* **435**, 207–211 (2005).
13. Schlich, R. & Axhausen, K. W. Habitual travel behaviour: Evidence from a six-week travel diary. *Transportation.* **30**, 13–36 (2003).
14. Eagle, N. & Pentland, A. Eigenbehaviours: identifying structure in routine. *Behav. Ecol. Sociobiol.* **63**, 1057–1066 (2009).
15. Klafter, J., Shlesinger, M. F. & Zumofen, G. Beyond Brownian motion. *Phys. Today* **49**, 33–39 (1996).
16. Gonzalez, M. C., Lind, P. G. & Herrmann, H. J. A system of mobile agents to model social networks. *Phys. Rev. Lett.* **96**, 088702 (2006).
17. Horner, M. W. & O-Kelly, M. E. S Embedding economies of scale concepts for hub networks design. *J. Transp. Geogr.* **9**, 255–265 (2001).
18. Kitamura, R., Chen, C., Pendyala, R. M. & Narayaran, R. Micro-simulation of daily activity-travelpatterns for travel demand forecasting. *Transportation* **27**, 25–51 (2000).
19. Colizza, V., Barrat, A., Barthelemy, M., Valleron, A.-J. & Vespignani, A. Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions. *PLoS Medicine* **4**, 95–110 (2007).
20. Eubank, S. *et al*. Controlling epidemics in realistic urban social networks. *Nature* **429**, 180–184 (2004).
21. Hufnagel, L., Brockmann, D. & Geisel, T. Forecast and control of epidemics in a globalized world. *Proc. Natl Acad. Sci. USA* **101**, 15124–15129 (2004).
22. Hedstrom, P. Experimental Macro Sociology: Predicting the Next Best Seller. *Science* **311**, 786–787. (2006).
23. Tumasjan, A., Sprenger, T. O., Sandner, P. G. & Welpe, I. M. Predicting elections with Twitter: What 140 characters reveal about political sentiment. *Proceedings of the fourth international aaai conference on weblogs and social media*. (2010).
24. Kirilenko, A. A., Kyle, A. S., Samadi, M. & Tuzun, T. The flash crash: The impact of high frequency trading on an electronic market. *Social Science Research Network*. (2011). http://dx.doi.org/10.2139/ssrn.1686004
25. Where's George? 2.4 – Track Your Dollar Bills, http://wheresgeorge.com website accessed July 30, 2013.
26. Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S. & Danforth, C. M. The Geography of Happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE* **8**(5), e64417 (2013).
27. Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A. & Danforth, C. M. Temporal Patterns of Happiness and Information in a Global-Scale Social Network: Hedonometrics and Twitter. *PLoS ONE* **6**(12), e26752 (2011).
28. Frank, M. R., Mitchell, L., Dodds, P. S. & Danforth, C. M. Online Appendix for "Happiness and the Patterns of Life: A Study of Geolocated Tweets", http://www.uvm.edu/storylab/share/papers/frank2013a website accessed July 30, 2013.
29. Zipf, G. Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology*. (1929).
30. Kloumann, I. M., Danforth, C. M., Harris, K. D., Bliss, C. A. & Dodds, P. S. Positivity of the English Language. *PLoS ONE* **7**(1), e29484 (2012).
31. Amazon's Mechanical Turk service. Available at https://www.mturk.com/ Accessed October 24, 2011.
32. Rand, D. G. The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *J Theor. Biol.* (2011).
33. Dodds, P. S. & Danforth, C. M. Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents. *Journal of Happiness Studies* (2009).
34. Aaron, S. & Joanna, B. Twitter Use 2012. *Technical report*, Pew Research Institute, 2012.
35. Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S. & Danforth, C. M. Online Appendix for "The Geography of Happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place." http://www.uvm.edu/storylab/share/papers/mitchell2013a website accessed July 30, 2013.
36. Sadilek, A., Kautz, H. & Silenzio, V. Modeling Spread of Disease from Social Interaction. *Sixth AAAI International Conference on Weblogs and Social Media (ICWSM)*. 2012.
37. Onnela, J.-P., Arbesman, S., Gonzlez, M. C., Barabsi, A.-L. & Christakis, N. A. Geographic Constraints on Social Network Groups. *PLoS ONE* **6**(4), e16939 (2011).
38. Bliss, C. A., Kloumann, I. M., Harris, K. D., Danforth, C. M. & Dodds, P. S. Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *Journal of Computational Science* **3**(5), 388–397 (2012).
39. U.S. Census Bureau Geography Division. *2010 Census TIGER/Line Shapefiles*. http://www.census.gov/geo/www/tiger/tgrshp2010/tgrshp2010.html, accessed February 2013.

Acknowledgements

For their helpful comments and input, we thank Brian Tivnan, James Bagrow, Yu-Ru Lin, Taylor Ricketts, Austin Troy, and Lisa Aultman-Hall. We are grateful for funding from the MITRE Corporation, the UVM Transportation Research Center, the Vermont Complex Systems Center, and the Vermont Advanced Computing Core, which was supported by NASA (NNX 08A096G). PSD was supported by NSF CAREER Grant No. 0846668.

Author contributions

C.M.D. and P.S.D. designed the research, M.R.F. prepared the figures, and C.M.D. and M.R.F. wrote the manuscript. M.R.F., L.M., P.S.D. and C.M.D. analyzed the data and reviewed the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Frank, M.R., Mitchell, L., Dodds, P.S. & Danforth, C.M. Happiness and the Patterns of Life: A Study of Geolocated Tweets. *Sci. Rep.* **3**, 2625; DOI:10.1038/srep02625 (2013).