arXiv:1101.5120v3 [physics.soc-ph] 11 Feb 2011

Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter

Peter Sheridan Dodds,[1,*] Kameron Decker Harris,[1] Isabel M. Kloumann,[1] Catherine A. Bliss,[1] and Christopher M. Danforth[1,†]

[1] *Department of Mathematics and Statistics, Center for Complex Systems, & the Vermont Advanced Computing Center, University of Vermont, Burlington, VT, 05401*

(Dated: February 14, 2011)

Individual happiness is a fundamental societal metric. Normally measured through self-report, happiness has often been indirectly characterized and overshadowed by more readily quantifiable economic indicators, such as gross domestic product. Here, we use a real-time, remote-sensing, non-invasive, text-based approach—a kind of hedonometer—to uncover collective dynamical patterns of happiness levels expressed by over 50 million users in the online, global social network Twitter. With a data set comprising nearly 2.8 billion expressions involving more than 28 billion words, we explore temporal variations in happiness, as well as information levels, over time scales of hours, days, and months. Among many observations, we find a steady global happiness level, evidence of universal weekly and daily patterns of happiness and information, and that happiness and information levels are generally uncorrelated. We also extract and analyse a collection of happiness and information trends based on keywords, showing them to be both sensible and informative, and in effect generating opinion polls without asking questions. Finally, we develop and employ a graphical method that reveals how individual words contribute to changes in average happiness between any two texts.

I. INTRODUCTION

One of the great modern scientific challenges we face lies in understanding macroscale sociotechnical phenomena—i.e., the behavior of decentralized, networked systems inextricably involving people, information, and machine algorithms—such as global economic crashes and the spreading of ideas and beliefs [1]. Description through quantitative measurement is essential to the advancement of any scientific field, and data transitions have revolutionized many areas [2–5] ranging from astronomy [6–8] to ecology and biology [9] to particle physics [10]. For the social sciences, the now widespread usage of the Internet has led to a collective, open recording of an enormous number of transactions, interactions, and expressions, marking a clear transition in our ability to quantitatively describe, and thereby potentially understand, previously hidden microscale mechanisms underlying sociotechnical systems.

While there are undoubtedly limits to that which may eventually be quantified regarding human behavior, recent studies have demonstrated a range of successful and practical methodologies, all impossible (if imaginable) prior to the Internet age. Three examples relevant to public health, markets, entertainment, history, evolution of language and culture, and prediction are (1) Google's digitization of over 15 million books and an initial analysis of the last two hundred years, showing language usage changes, censorship, dynamics of fame, and time compression of collective memory [11, 12]; (2) Google's Flu Trends [13–15] which allows for real-time monitoring of flu outbreaks through the proxy of user search; and (3) the highly accurate prediction of box office success based on the rate of online mentions of individual movies [16].

Out of the many possibilities in the 'big data' age of social sciences, we focus here on the measuring, describing, and understanding of the well-being of large populations. A measure of 'societal happiness' is a crucial adjunct to traditional economic measures such as gross domestic product and of fundamental scientific interest in its own right [17–19]. Our specific overall objective is to use web-scale text analysis [20] to remotely sense societal-scale levels of happiness using the singular source of the microblog and social networking service Twitter. In essence, our method for measuring the happiness of a given text, which we describe fully below and which we introduced in [20], entails word frequency distributions combined with independently assessed, numerical estimates of the 'happiness' of words [21]. Twitter is simple in nature, allowing users to place brief, text-only expressions online—'status updates' or 'tweets'—that are no more than 140 characters in length. As we will show, Twitter's framing tends to yield in-the-moment expressions that reflect users' current experiences, making the service an ideal candidate input signal for a real time societal 'hedonometer' [22].

There is an important psychological distinction between an individual's current, experiential happiness and their longer term, reflective evaluation of their lives [23], and in using Twitter, our approach is most obviously tuned to the former kind. Nevertheless, by following the written expressions of individual users over long time periods, we are potentially able to infer details of happiness dynamics such as individual stability, social correlation and contagion [24], and connections to well-being and health [17, 23].

*Electronic address: peter.dodds@uvm.edu
†Electronic address: chris.danforth@uvm.edu

This paper is the first in a series where we explore happiness as a function of time, space, demographics, and network structure using Twitter as a data source. Here, we focus on our essential findings regarding temporal variations in happiness including: the overall time series; regular cycles at the scale of days and weeks; time series for subsets of tweets containing specific keywords (e.g., ‘Obama’); and detailed comparisons between texts at the level of individual words. We also compare happiness levels with measures of information content. For information, as we explain in detail below, we employ an estimate of lexical size (or effective vocabulary size) which is related to species diversity for ecological populations and is derived from generalized entropy measures [25].

Our findings and approach complement a number of related efforts undertaken in recent years regarding happiness and well-being including large-scale surveys carried out by Gallup [26]; population-level happiness measurements carried out by Facebook’s internal data team [27]; work focusing directly on sentiment detection based on Twitter [28–31]; and survey-based, psychological profiles as a function of location, such as for the United States [32].

We structure our paper as follows: in Secs. II and III, we detail our methods for measuring happiness and information content; in Sec. IV, we describe our data set; in Sec. V, we present and discuss the overall time series for happiness and information; in Secs. VI and VII we examine the average weekly and daily cycles in detail; in Sec. VIII, we explore happiness and information time series for tweets containing keywords and text elements; and in Sec. IX, we offer some concluding remarks.

II. MEASURING HAPPINESS

Our simple, fast method for measuring the happiness of texts hinges on two key components [20]: (1) human evaluations of the happiness of a set of individual words, described below; and (2) a naive algorithm for scaling up from individual words to texts. For the algorithm, we first use a straightforward pattern-matching script to extract the frequency of individual words in a given text T. We then compute the weighted average level of happiness for the text as

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

where f_i is the frequency of the ith of N distinct words w_i for which we have an estimate of average happiness, $h_{\rm avg}(w_i)$, and $p_i = f_i / \sum_{j=1}^{N} f_j$ is the corresponding normalized frequency.

As for the happiness of individual words, we capitalize on results from the 1999 Affective Norms for English Words (ANEW) study by Bradley and Lang [21]. In the ANEW study, participants were presented with isolated individual words and asked to grade them on an unhappy-happy integer scale ranging from 1 to 9 (levels of excitement and dominance were also surveyed). Participants were informed that a score of 9 should correspond to them feeling completely “happy, pleased, satisfied, contented, hopeful,” and to record a lowest score of 1 if they felt “completely unhappy, annoyed, unsatisfied, melancholic, despaired, or bored.” The ANEW study comprised 1034 words which were broadly chosen for their emotional and meaningful import. We use the average happiness scores as reported for each word by Bradley and Lang. The average happiness of words were distributed across the 1–9 range, with some illustrative examples being $h_{\rm avg}$(love) = 8.72, $h_{\rm avg}$(food) = 7.65, $h_{\rm avg}$(reunion) = 6.48, $h_{\rm avg}$(truck) = 5.47, $h_{\rm avg}$(vanity) = 4.30, $h_{\rm avg}$(greed) = 3.51, $h_{\rm avg}$(hate) = 2.12, and $h_{\rm avg}$(funeral) = 1.39. As this short list indicates, the evaluations are sensible with neutral words averaging around 5 [e.g., $h_{\rm avg}$(barrel) = 5.05, $h_{\rm avg}$(chair) = 5.08].

Text:	$h_{\rm avg}$	Words with a similar score:
Soul/Gospel music lyrics [20]	6.9	chocolate (6.88), leisurely (6.88), penthouse (6.81)
Pop music lyrics [20]	6.7	dream (6.73), honey (6.73), sugar (6.74)
Dante’s Paradise [33]	6.5	muffin (6.57), rabbit (6.57), smooth (6.58)
Tweets, 9/9/2008 to 12/31/2010 (present work)	6.4	thought (6.39), face (6.39), blond (6.42)
Rock music lyrics [20]	6.3	church (6.28), tree (6.32), air (6.34)
Enron Emails [34]	6.2	clouds (6.18), alert (6.20), computer (6.24)
State of the Union Messages [20]	6.1	grass (6.12), idol (6.12), bottle (6.15)
New York Times (1987–2007) [35]	6.0	hotel (6.00), tennis (6.02), wonder (6.03)
Blogs [20]	5.8	owl (5.80), whistle (5.81), humble (5.86)
Dante’s Inferno [33]	5.5	glacier (5.50), repentant (5.53), mischief (5.57)
Metal/Industrial music lyrics [20]	5.4	lamp (5.41), elevator (5.44), truck (5.47)

TABLE I: Average happiness $h_{\rm avg}$ for some example texts, to provide a sense of how our instrument works. Our measurement technique builds on happiness scores for individual words that range from 1 to 9 (see Sec. II), and for nearly all texts we have analyzed we find $5 \le h_{\rm avg} \le 7$.

We find that our measure typically places average happiness for texts between 5 and 7. To give the reader a better sense of our instrument’s scale, we show in Tab. I how a sample of texts are located on the average happiness spectrum. In the ANEW study, the unhappy-happy scale was reported as psychological valence, or simply valence, a standard terminology [36] which we followed in our first work; here we use the more straightforward nomenclature of 'happiness.'

We address several key aspects and limitations of our measurement. First, as with any sentiment analysis technique, our instrument is fallible for smaller texts, especially at the scale of a typical sentence, where ambiguity may render even human readers unable to judge meaning or tone [37]. Nevertheless, problems with small texts are not our concern, as our interest here is in dealing with and benefiting from very large data sets. Second, we are also effectively extracting a happiness level as perceived by a generic reader who sees only word frequency. Indeed, our method is purposefully more simplistic than traditional natural language processing (NLP) algorithms which attempt to infer meaning (e.g., OpinionFinder [38, 39]) but suffer from a degree of inscrutability. By ignoring the structure of a text, we are of course omitting a great deal of content; nevertheless, we have shown using bootstrap-like approaches that our method is sufficiently robust as to be meaningful for large enough texts [20]. Third, we quantify only how people appear to others; as should be obvious, our method cannot divine the internal emotional states of specific individuals or populations. In attempting to truly understand a social system's potential dynamical evolution, we would have to account for hidden but accessible internal ranges and states of emotions, beliefs, etc. However, a person's exhibited emotional tone, now increasingly filtered through the signal-limiting medium of written interactions (e.g., status updates, emails, and text messages), is that which other people evidently observe and react to. Last, by using a simple kind of text analysis, we are able to non-invasively, remotely sense the happiness of very large numbers of people via their written, open, web-scale output. Crucially, we do not ask people how happy they are, we merely observe how they behave online. As such, we avoid the many difficulties associated with self-report [40–42]. We refer the reader to our initial work for a more in depth discussion of our measurement technique [20].

III. MEASURING WORD DIVERSITY

In quantifying a text's information content, we use concepts traditionally employed for estimating species diversity in ecological studies [25] which build on information theoretic approaches. As we outline below, direct measures of information can be transformed into estimates of lexical size (or word diversity), with the benefit that comparisons of the latter are more readily interpretable.

A first observation is that the sheer number of distinct words is not a useful representation of lexical size; because natural texts generally exhibit highly skewed distributions of word frequencies, such a measure discards much salient information. To arrive at a more useful quantity, we start with direct measures of information. Two specific examples of use here are Simpson's concentration [43] and Shannon's entropy [44]. For a given text, all information measures depend on the ith distinct word's normalized frequency of occurrence p_i. Simpson's concentration is given by $S = \sum_i p_i^2$, the probability that any two words chosen at random will be the same. Simpson's concentration is related to the Gini coefficient G, which is often used to characterize income inequality, as $S = 1 - G$; for text analysis, G represents the probability that two randomly chosen words are different. Shannon's entropy is defined as $H = -\sum_i p_i \ln p_i$, and is proportional to the average number of bits required per word to efficiently encode the overall text. Moving beyond these two measures, we also consider the generalized entropy $J_q = \sum_i p_i^q$ (note that $S = J_2$). In varying q, we tune the relative importance of common versus rare words, with large q favoring common ones.

One drawback of these information measures is that their values can be hard to immediately interpret. To make comparisons between the information content of texts more understandable, if by adding an extra step, we use these information measures to compute an equivalent lexical size, $N_q^{\rm eq}$. This size is the number of words that would yield the same information measure if all words appeared with equal frequency [25]. Simple calculations give $N_2^{\rm eq} = 1/S$ and $N_1^{\rm eq} = e^H$. For generalized entropy, $N_q^{\rm eq} = J_q^{1/(1-q)}$ if $q \neq 1$ and $\lim_{q\to 1} N_q^{\rm eq} = e^{-\sum_i p_i \ln p_i} = e^H$. (Thus, in terms of lexical size estimates, generalized entropy naturally corresponds to Shannon entropy when $q \to 1$, even though $H \neq J_1 = 1$.)

We observe that the lexical sizes $N_q^{\rm eq}$ for $q \geq 1$ closely follow the same trends for the data we analyse here. In therefore needing to show only one representative measure among the $N_q^{\rm eq}$, we choose $N_{\rm S} = N_2^{\rm eq}$ based on Simpson's concentration S because it holds several theoretical and practical benefits: (1) S has the natural probablistic interpretation given above; (2) The quantity p_i^2 decays sufficiently rapidly that we need not be concerned about subsampling heavy tailed distributions (see methods, Sec. IX); (3) In comparing two texts, the contributions to $N_{\rm S}$ due to changes in individual word frequencies combine linearly and thus can be easily ranked (see the last part of Sec. VI). From here on, we will focus on $N_{\rm S}$ which we will refer to as a text's 'Simpson lexical size.'

In wanting to rapidly compare in detail the Simpson lexical size of many pairs of massive texts assembled on the fly (e.g., by finding all tweets that contain a particular keyword), a computational difficulty arises in the generation and cross-comparison of word frequencies. Indeed, any word-based information comparison suffers from the same problem. To facilitate such comparisons, we took the top 50,000 words from a large part of the overall Twitter corpus (see methods, Sec. IX), as a standardized list. Texts could then be transformed into vectors of word frequencies with a fixed ordering of words across all texts. The number 50,000 was chosen both for computational ease and the fact that Simpson's lexical size could reliably be computed without concerns over convergence as mentioned above.

Tweet attributes:
Tweet text
Unique tweet ID
Date and time tweet was posted†
UTC offset (from GMT)
User's location
User ID
Date and time user's account was created
User's current follower count
User's current friends count
User's total number of tweets
In-reply-to tweet ID*
In-reply-to user ID*
Retweet (Y/N)

TABLE II: A list of key informational attributes accompanying each tweet. Information regarding the time of posting was altered (†) on May 21, 2009 so that local time rather than Greenwich Mean Time (GMT) was reported. If a tweet is a reply to a previous tweet, the attributes also include those indicated by an asterisk: the ID of the specific tweet's and user's ID. Twitter initially issued tweets in xml format before moving to the json standard [46].

IV. DESCRIPTION OF DATA SET

Since its inception, Twitter has provided various kinds of dedicated data feeds for research purposes. For the results we present here, we collected tweets over a twenty-eight month period running from September 9, 2008 to December 31, 2010. To the nearest million, our data set comprises 28.446 billion words contained in 2.772 billion tweets posted by over 50 million individual users. Our collection represents 8% of all tweets posted up to December 31, 2010 (based on number of tweets up until November 6, 2010) [45]. In the full data set, we identified 1.046 billion ANEW study words, representing approximately 3.7% of all words, typical of other texts we have analysed such as blogs, books, and State of the Union Addresses [20].

Our rate of gathering tweets was not constant over time, with regions of stability connected by short periods of considerable flux (shown later in Fig. 2C). These changes were due to periodic alterations in Twitter's feed mechanism as the company adjusted to increasing demand on their service [46]. Twitter's tremendous growth in usage and importance over this time frame lead to several service outages, and generated considerable technical issues for us in handling and storing tweets. Nevertheless, we were able to amass a very large data set, particularly so for one in the realm of social phenomena. Our collection rate was at most on the order of 10% of all tweets. By December 31, 2010, were receiving roughly 10 million tweets per day (roughly 7000 per minute), and there were only a few days for which we did not record any data.

Each tweet delivered by Twitter was accompanied by a basic set of informational attributes; we list the salient ones in Tab. II, and summarize them briefly here. First, for all tweets, we have a time stamp referring to a single world clock running on US Eastern Standard Time; and from May 21, 2009 onwards, we also have local time as well. User location is available for some tweets and varies from current latitude and longitude, as reported for example by a smartphone, to a static, free text entry of a home city along with state and country. For measures of social interactions, we have a user's current follower and friend counts (but no information on who the followers and friends are), and if a tweet is made in reply to another tweet, we also have the identifying number (ID) of the latter. Finally, a 'retweet' flag ('RT') indicates if a tweet is a rebroadcasting of another tweet, encoding an important kind of information spreading in the Twitter network.

Against the many benefits of using a data source such as Twitter, there are a number of reasonable concerns to be raised, notably representativeness. First, in terms of basic sampling, tweets allocated to data feeds by Twitter were effectively chosen at random from all tweets. Our observation of this apparent absence of bias in no way dismisses the far stronger issue that the full collection of tweets is a non-uniform subsampling of all utterances made by a non-representative subpopulation of all people. While the demographic profile of individual Twitter users does not match that of, say, the United States, where the majority of users currently reside [47], our interest is in finding suggestions of universal patterns. Moreover, we note that like many other social network services, Twitter accommodates organizations as users, particularly news services. Twitter's user population is therefore a blend of individuals, groups of individuals and automated services such as bots [48], representing an emerging kind of disaggregated, crowd-sourced media. Thus, rather than analysing signals from a few news outlets, which in theory represent and reflect the opinions and experiences of many, we now have access to signals coming directly from a vast number of individuals. Moreover, in our treatment, tweets from, say, the New York Times or the White House are given equal weight to those of any person-on-the-street.

In sum, we see two main arguments for pursuing the massive data stream of Twitter: (1) the potential for describing universal human patterns, whether they be emotional, social, or otherwise; and (2) the current importance of Twitter across all of media [49] (surprising as that may be to critics).

A preliminary glance at the data set shows that the raw word content of tweets does appear to reflect people's current circumstances. For example, Fig. 1 shows normalized daily frequencies for two food-based sets of words, binned by hour of the day. Fig. 1A shows that, as we would expect, the words 'breakfast', 'lunch', and 'dinner' respectively peak during the hours 8–9 am, 12–1 pm, and 6–7 pm. In Fig. 1B, we observe that the words 'starving', 'chicken' 'hungry', 'eat', and 'food', all following a similar cycle with three relative peaks, one around midday, a smaller one before dinner, and another

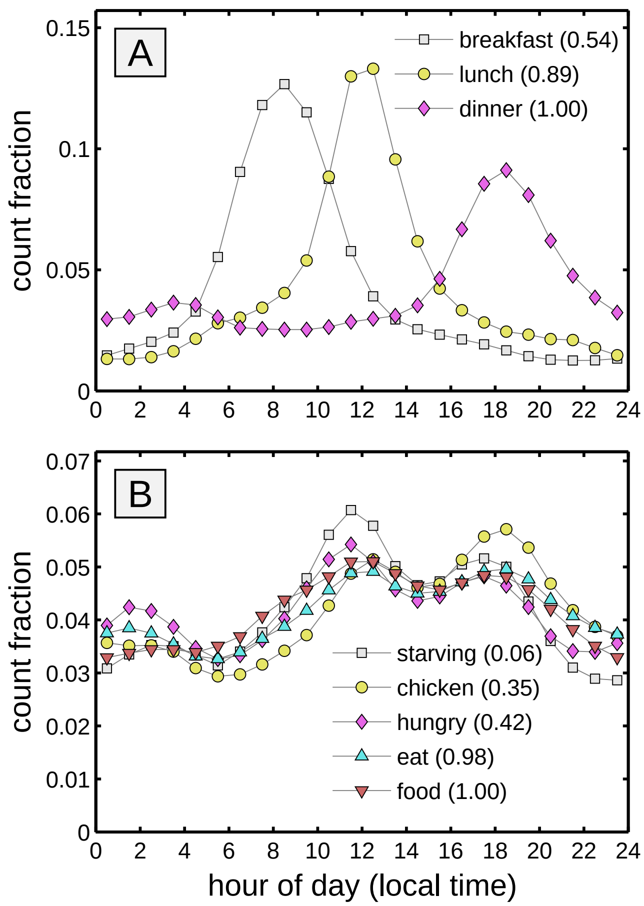

FIG. 1: Daily trends for example sets of commonplace words appearing in tweets. For purposes of comparison, each curve is normalized so that the count fraction represents the fraction of times a word is mentioned in a given hour relative to a day. The numbers in parentheses indicate the relative overall abundance normalized for each set of words by the most common word. Data for these plots is drawn from approximately 26.5 billion words collected from May 21, 2009 to December 31, 2010 inclusive, with the time of day adjusted to local time by Twitter from the former date onwards. The words 'food' and 'dinner' appeared a total of 2,994,745 (0.011%) and 4,486,379 (0.016%) times respectively.

in the early morning. These trends suggests more generally that words that are correlated conceptually will be similarly congruent in their temporal patterns in tweets. Other quotidian words follow equally reasonable trends: the word 'sunrise' peaks between 6 and 7 am, while 'sunset' is most prominent around 6 pm; and the daily high for 'coffee' occurs between 8 and 9 am. Regular cultural events also leave their imprint with two examples from television being 'lost' (for the show 'Lost') and 'idol' (for 'American Idol') both sharply maximizing around their airing times in the evening. Thus, while not statistically exhaustive, we have reassuring, commonsensical support for the in-the-moment nature of tweets, and we move on to our main descriptive focus: patterns of happiness.

V. OVERALL TIME DYNAMICS OF HAPPINESS AND INFORMATION

We observe a variety of temporal trends in happiness and information content across time scales of hours, days, and months. In Fig. 2A we present the average happiness time series with tweets binned by date. The accompanying plots, Figs. 2B and 2C, show Simpson lexical size $N_{\rm S}$ and the number of ANEW study words contained in our tweets over the same time period. We expect averaging over all but time to leave only truly system wide signals, and as we show later in Section VIII, subsets of tweets exhibit markedly different temporal trends.

At the level of months, we see that average happiness gradually increased over the last months of 2008 and 2009, and dropped in January of both 2009 and 2010. Apart from another gradual upward trend that ran from January to April, 2009, the overall time series was otherwise reasonably stable. Moving down to timescales less than a month, we see a clear weekly signal. Generally, the peak occurred over the weekend, and the nadir on Monday and Tuesday. We return to and examine the weekly cycle in detail in the following section.

At the scale of a day, we find a number of dates which strongly deviate in their happiness levels from nearby dates, and we indicate these by filled circles and icons in Fig. 2A. In both 2008 and 2009, Christmas Day returned the highest levels of happiness, followed by Christmas Eve. Other notable dates include New Year's Eve and Day, Valentine's Day, Thanksgiving, Fourth of July, Easter Sunday, Mother's Day, Father's Day, and the day of Michael Jackson's death. The strength of Thanksgiving and Fourth of July reflects the fact that while Twitter is a global service, the majority of users still come from the United States [47], an effect magnified further by our present measure of happiness using only English words.

Michael Jackson's death is both the only date for which we see a substantial, system-wide negative drop in happiness, and the only singular event to register notably in the overall time series—all other dates mark well established religious, cultural, and/or national annual events. There are several other events that had lesser overall impact but can still be discerned as drops. These include the death of actor Patrick Swayze (September 14, 2009) and the premiere and finale of the last season of the highly rated television show 'Lost', marked by a drops in our time series at February 2 and May 24, 2010, and largely due to the word 'lost' having a low happiness score of $h_{\rm avg}$=2.82.

A number of these departures for specific dates qualitatively match observations we made in our earlier work on blogs [20], though we make any comparison tentatively as for blogs we focused on sentences written in the first person containing a conjugation of the verb 'to feel' [50]. For example, Christmas Eve and Day, New Year's Eve and Day, and Valentine's Day all exhibit jumps in happiness in both tweets and 'I feel...' blog sentences. Both

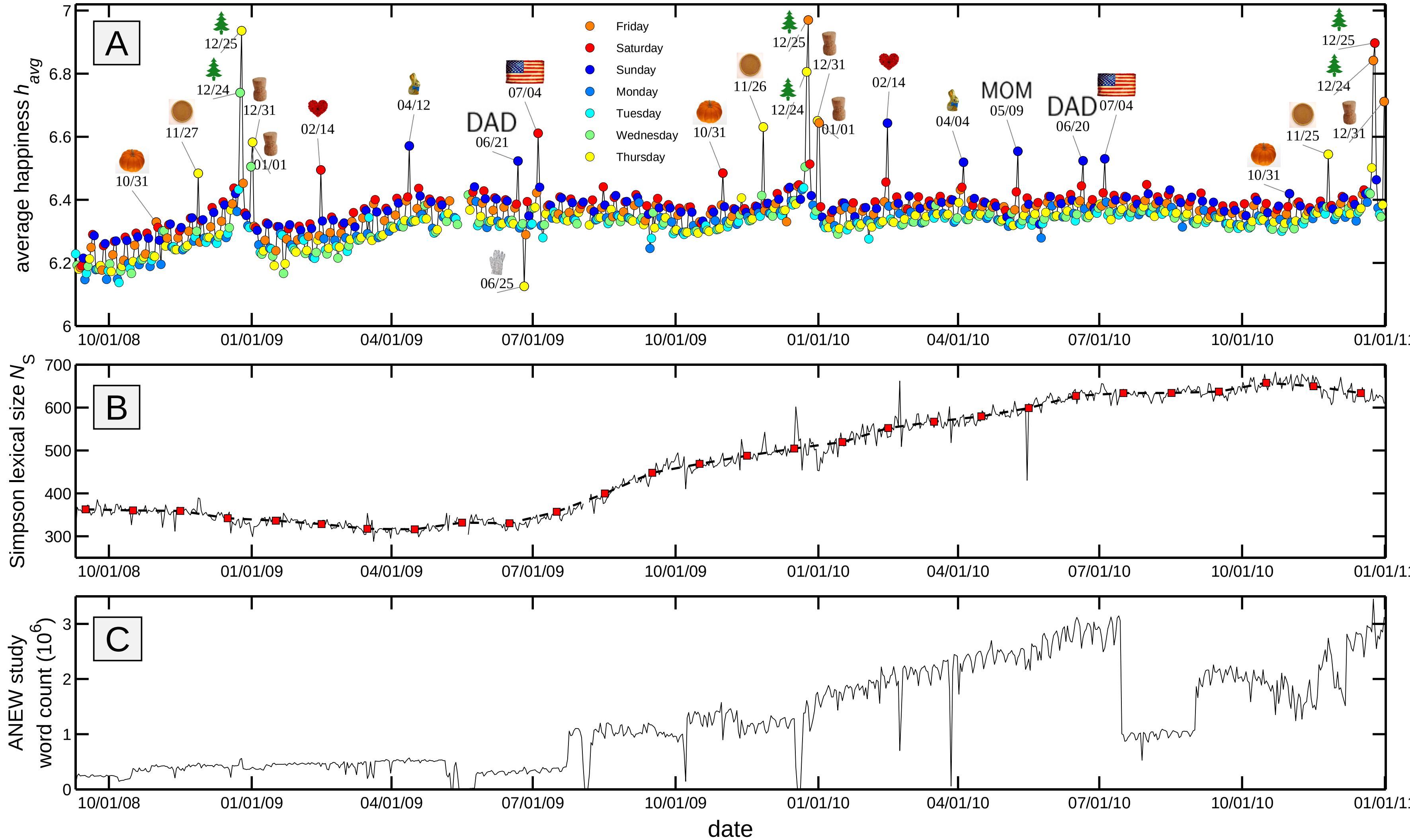

FIG. 2: **A.** Overall happiness time series for all tweets averaged by individual day (zoomable online; see Sec. III for measurement explanation). Circle colors indicate day of the week as per the legend; a regular weekly cycle is clear with the red and blue of Saturday and Sunday typically the high point (examined further in Fig. 3). The dates on the horizontal axis indicate the start of the month. We label outlier dates for which average happiness strongly departs from the local pattern (we include annual dates for which at least one example could be considered an outlier; see text). Post May 21, 2009, we use reported local time to assign tweets to particular dates. **B.** Simpson lexical size N_S as a function of date using Simpson's concentration as the base entropy measure (solid gray line; see Sec. III). The red squares with the dashed line show N_S as a function of calendar month. **C.** The number of ANEW study words extracted from all tweets as a function of date [20, 21]. The variability was due to planned changes and irregularities in Twitter's data feed coupled with technical problems in collecting such large amounts of data. For both the happiness and Simpson lexical size plots, we omit dates for which we have less than 1000 ANEW study words.

time series also show a pronounced drop for Michael Jackson's death. However, tweets did not register a similar lift as blogs for the US Presidential Election in 2008 and Inauguration Day, 2009, while positive sentiment for both Mother's and Father's Day, the Fourth of July, are much more evident in tweets. Lastly, blogs typically showed drops for September 10 and/or 11 that are absent in tweets.

For information content over this same time period (Fig. 2B), we see a strong increase in Simpson lexical size N_{S} climbing from approximately 300 to 700 words beginning around July, 2009. (For $q \geq 1$, generalized word diversities all follow the same trajectory with N_q^{eq} increasing as q decreases.) We also indicate in the same plot N_{S} measured at the scale of months (red squares). The smoothness of the resulting curve shows that N_{S} is unaffected by the two issues of missing data and non-uniform sampling rates. (Note that the month estimates of N_{S} are computed from the word distribution for the month and are not simply averages of daily values of N_{S}.)

By examining shifts in word usage, we are able to attribute the more than doubling of N_{S} to a strong relative increase in non-English languages, notwithstanding the dramatic growth in English language tweets. Recalling that the most common words such as articles and prepositions figure most strongly in the computation of the Simpson word diversity, we see the dominant growth in Spanish ('que', 'la', 'y','en', 'el'); a few other example languages making headway are Portuguese ('pra'), and Indonesian ('yg'). By contrast, English words appear relatively less (including the word 'twitter') while a minority of words move against the general diversification by appearing more frequently, with prominent examples being the abbreviations 'RT' (for retweet) and 'lol' (for laugh out loud).

VI. WEEKLY CYCLE

As we saw in Fig. 2A, a pronounced weekly cycle is present in the overall time series. To reveal this feature more clearly, we compute average happiness h_{avg} as a function of day of the week, Fig. 3. Taking tweets for which we have local time information (May 21, 2009 onward), we show two curves, one for which we include all data (crosses), and one for which we exclude the outlier days we identified in Fig. 2A (labelled dates accompanied by icons). Including outlier days yields a higher average happiness, and the difference between the two curves is most pronounced on Thursday with minor increases on Friday, Saturday, and Sunday. These discrepancies are explained by Thanksgiving in 2009 and 2010, and Christmas Eve and Day and New Year's Eve and Day falling on Thursday and Friday in 2009 and Friday and Saturday in 2010.

We take the reasonable step of focusing on the data with outlier days removed. We see Saturday has the highest average happiness ($h_{\mathrm{avg}} \simeq 6.40$), closely followed by

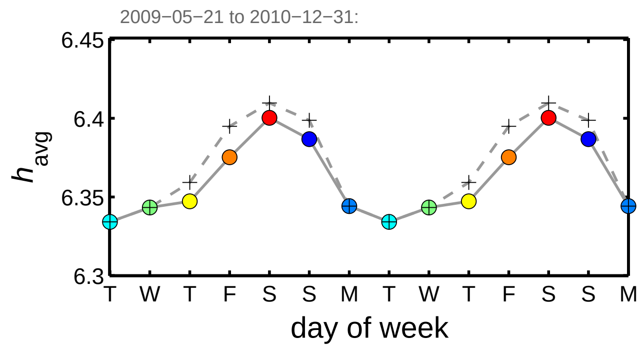

FIG. 3: Average happiness as a function of day of the week for our complete data set. To make the average weekly cycle more clear, we repeat the pattern for a second week. The crosses indicate happiness scores based on all data, while the filled circles show the results of removing the outlier days indicated in Fig. 2A. The colors for the days of the week match those used in Fig. 2A. To circumvent the non-uniform sampling of tweets throughout time, we compute an average of averages: for example, we find the average happiness for each Monday separately, and then average over these values, thereby giving equal weight to each Monday's score. We use only data from May 21, 2009 on, for which we have a local timestamp.

Sunday and then a large drop to Monday's level. The weekly low occurs on Tuesday, followed by small increases on both Wednesday and Thursday ($h_{\mathrm{avg}} \simeq 6.33$). We see a jump on Friday, the third happiest day of the week, leading back to the peak of Saturday. Roughly similar patterns have been found in Gallup polls [26], Facebook by the company's internal research team [27], and in analyses of smaller collections of tweets [51]. (In the last work and in contrast to our findings here for a data set tenfold larger in size, Thursday evening was identified as the low point of the week.) The weekend peak in the cycle conforms with everyday intuition. The shallow minimum on Tuesday goes against standard notions of the Monday blues with its back-to-work nature, and Wednesday's middle-of-the-week labelling as the work week's hump day. In our earlier work on blogs, we saw a statistically significant but much weaker cycle for the days of the week; the high and low days were Sunday ($h_{\mathrm{avg}} = 5.856$) and Wednesday ($h_{\mathrm{avg}} = 5.842$), a difference of only 0.014 [20], around 17% of that for tweets. The discrepancy appears to be due to the in-the-moment character of Twitter versus the reflective one of blogs.

With any observed pattern, a fundamental issue is universality. Is the three day midweek low followed by a peak around Saturday a pattern we always see, given enough data? Further inspection of our Twitter data set shows a constancy in the weekly cycle occurring over time. In Fig. 4, we aggregate tweets for days of the week for four time ranges, approximately equal in duration. As before, we show the weekly pattern for all days (crosses, dashed curve) and with outlier days marked in Fig. 2A removed (disks, solid curve). The major differences we observe between these two curves in the four panels are predom-

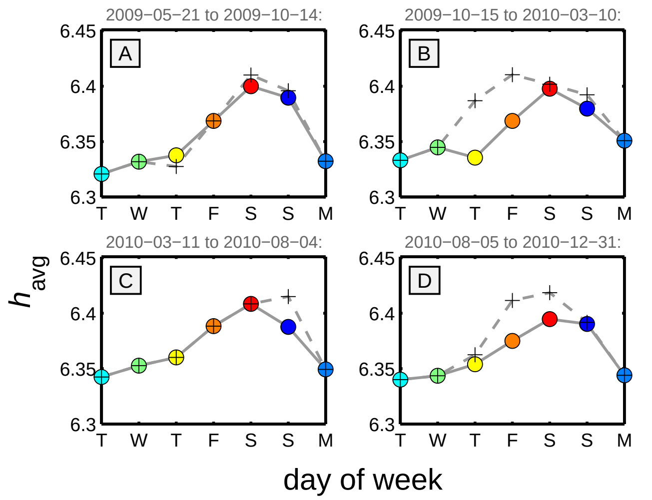

FIG. 4: Average of daily average happiness for days of the week over four consecutive time periods of approximately five months duration each. As per Fig. 3, crosses are based on all days, circles for days excluding outlier days marked in Fig. 2. To facilitate comparison, the vertical scale is the same in each plot and matches that used in Fig. 3.

inantly explained as before by Christmas, New Year's, and Thanksgiving. In terms of universality, we again see that Friday-Saturday-Sunday represents the peak while Tuesday's level is the minimum in each period. Only for Thursday in Fig. 4B, do we see a change in the overall ordering of days. Thus, we have some confidence that the overall weekly cycle of happiness shown in Fig. 3 is a fair description of what appears to be a robust pattern of users' expressed happiness.

When making a cross-comparison of two or more texts via a single summary statistic, such as average happiness and lexical size, we naturally need to look further into why a given measure shows variation. In Fig. 5, we provide a 'word shift graph' for Saturdays ($h_{\rm avg} = 6.40$) relative to Tuesdays ($h_{\rm avg} = 6.33$). We will use these graphs to illuminate how the differences between two texts' happiness levels arise from changes in underlying word frequency, and we explain them here in detail.

Consider two texts $T_{\rm ref}$ (for reference) and $T_{\rm comp}$ (for comparison) with happiness scores $h_{\rm avg}^{\rm (ref)}$ and $h_{\rm avg}^{\rm (comp)}$. If we wish to compare $T_{\rm comp}$ relative to $T_{\rm ref}$ then, using Eq. (1), we can write

$$h_{\rm avg}^{\rm (comp)} - h_{\rm avg}^{\rm (ref)} = \sum_{i=1}^{N} h_{\rm avg}(w_i) \left[p_i^{\rm (comp)} - p_i^{\rm (ref)} \right]$$
$$= \sum_{i=1}^{N} \left[h_{\rm avg}(w_i) - h_{\rm avg}^{\rm (ref)} \right] \left[p_i^{\rm (comp)} - p_i^{\rm (ref)} \right] \qquad (2)$$

where we have employed the fact that

$$\sum_{i=1}^{N} h_{\rm avg}^{\rm (ref)} \left[p_i^{\rm (comp)} - p_i^{\rm (ref)} \right] = h_{\rm avg}^{\rm (ref)} \sum_{i=1}^{N} \left[p_i^{\rm (comp)} - p_i^{\rm (ref)} \right]$$
$$= h_{\rm avg}^{\rm (ref)} (1-1) = 0.$$

In introducing the term $-h_{\rm avg}^{\rm (ref)}$, we are now able to make clear the contribution of the ith word to the difference $h_{\rm avg}^{\rm (comp)} - h_{\rm avg}^{\rm (ref)}$. From the form of Eq. (2), we see that we need to consider two aspects in determining the sign of the ith word's contribution:

1. whether or not the ith word is on average happier than text $T_{\rm ref}$'s average, $h_{\rm avg}^{\rm (ref)}$; and
2. whether or not the ith word is relatively more abundant in text $T_{\rm comp}$ than in text $T_{\rm ref}$.

We will signify a word's happiness relative to text $T_{\rm ref}$ by + (more happy) and − (less happy), and its relative abundance in text $T_{\rm comp}$ versus text $T_{\rm ref}$ with ↑ (more prevalent) and ↓ (less prevalent). Combining these two binary possibilities leads to four cases:

+↑: if a word is happier than text $T_{\rm ref}$ (+) and appears relatively more often in text $T_{\rm comp}$ (↑), it will constitute a positive contribution to the overall difference $h_{\rm avg}^{\rm (comp)} - h_{\rm avg}^{\rm (ref)}$;

−↓: if a word is less happy than text $T_{\rm ref}$ (−) and appears relatively less in text $T_{\rm comp}$ (↓), this will also lead to a positive gain;

+↓: a word that is happier than text $T_{\rm ref}$ (+) but less abundant in text $T_{\rm comp}$ (↓), leading to a decrease in average happiness; or

−↑: a word that is less happy than text $T_{\rm ref}$ (−) while more abundant in text $T_{\rm comp}$ (↑), also leading to a drop in average happiness.

For the convenience of visualization, we normalize the summands in Eq. (2) and convert to percentages to obtain:

$$\delta h_{\rm avg,i} = \frac{100}{\left| h_{\rm avg}^{\rm (comp)} - h_{\rm avg}^{\rm (ref)} \right|} \underbrace{\left[h_{\rm avg}(w_i) - h_{\rm avg}^{\rm (ref)} \right]}_{+/-} \underbrace{\left[p_i^{\rm (comp)} - p_i^{\rm (ref)} \right]}_{\uparrow/\downarrow}, \qquad (3)$$

where $\sum_i \delta h_{\rm avg,i} = \pm 100$, depending on the sign of the difference in happiness between the two texts, $h_{\rm avg}^{\rm (comp)} - h_{\rm avg}^{\rm (ref)}$, and where we have indicated the terms to which the symbols +/− and ↑ / ↓ apply. We call $\delta h_{\rm avg,i}$ the per word happiness shift of the ith word.

With these definitions in hand, we return to explaining Fig. 5, the word shift graph comparing tweets made on Saturdays relative to those made on Tuesdays. The bar graph shows the first 50 words ranked by their absolute contribution to the change in average happiness of Saturdays over Tuesdays, $|\delta h_{\rm avg,i}|$. We see examples of each of the four ways words can contribute. The two kinds of positive changes dominate with 41 of the top 50 changes, including more of 'love', 'party', 'fun', 'happy', 'wedding', 'movie', and 'beach' (all +↑), and less of

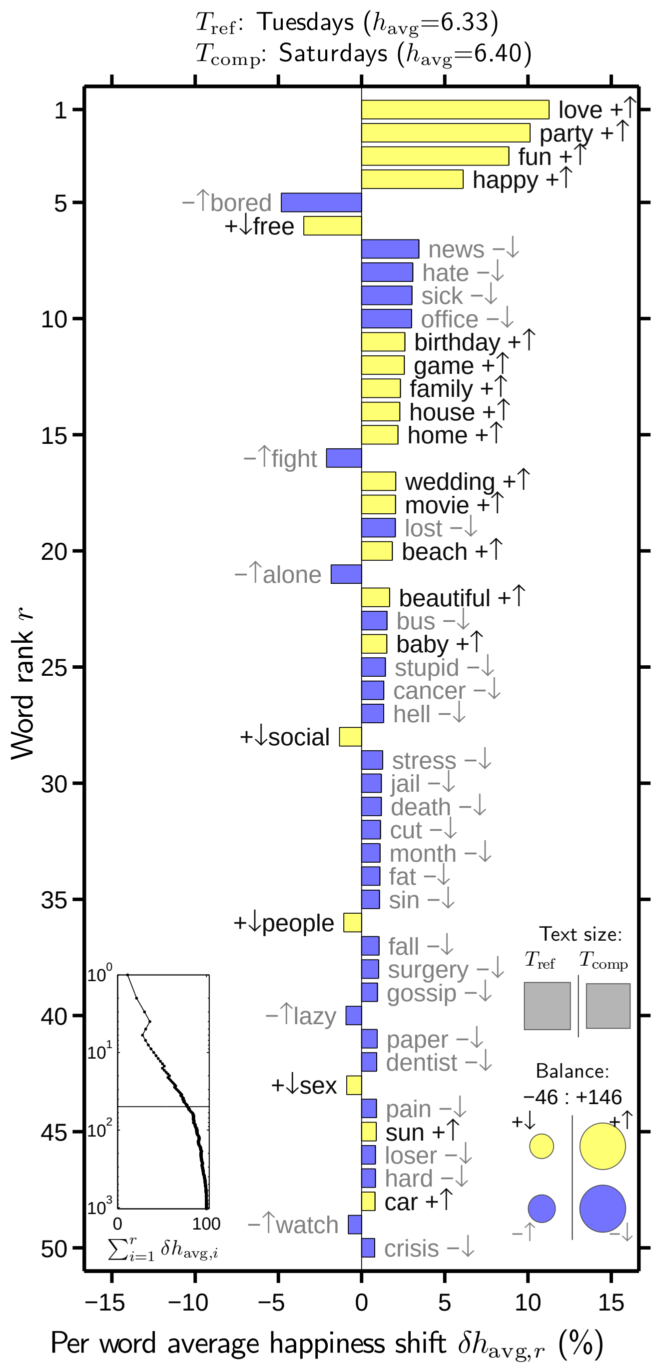

FIG. 5: Word shift graph showing how changes in word frequencies give Saturdays a higher overall happiness than Tuesdays. Words are ranked by their percentage contribution to the change in average happiness, $\delta h_{\mathrm{avg,i}}$. Tuesdays are set as the reference text (T_{ref}) and Saturdays as the comparison text (T_{comp}). How individual words contribute to the shift is indicated by two symbols: $+/-$ shows the word is more/less happy than Tuesdays as a whole, and $\uparrow$ / $\downarrow$ shows that the word is more/less relatively prevalent on Saturdays than on Tuesdays. Black and gray font additionally encode the $+$ and $-$ distinction respectively. The left inset panel shows how the ranked 1034 ANEW study words [20, 21] combine in sum (word rank r is shown on a log scale). The four circles in the 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. See Eqs. 2 and 3 and surrounding text for complete details.

'news', 'hate', 'sick, 'office', 'stress', and 'dentist' (all $-\downarrow$). These changes are readily interpretable, with the weekend involving more leisure and family time, and a relative absence of work and work-related concerns. Words in the top 50 which move against the general trend are the more prevalent, relatively negative words 'bored', 'fight', 'alone' and 'lazy' ($-\uparrow$), and the less frequent positive words 'free', 'social', and 'people' ($+\downarrow$). Thus while Saturdays may be on average happier than Tuesdays, we see predictable elements of boredom and loneliness appearing as well.

The three insets of Fig. 5 expand the story provided by the main figure in the following ways. First and simplest is the pair of gray squares on the right which show, by their area, the relative sizes of the two texts, as measured by the total number of ANEW study words (the absolute number of words is not indicated). For the example of Fig. 5, we see that we have slightly more ANEW study words for Tuesdays (T_{ref}) than Saturdays (T_{comp}).

Second, on the bottom left, the graph shows the cumulative sum of the individual word contributions, $\sum_{i=1}^{r} \delta h_{\mathrm{avg,i}}$ as a function of $\log_{10} r$ where r is word rank. The graph shows how rapidly the word contributions converge to $\pm 100\%$ as we include all 1034 ANEW study words. The solid line marks 50 words, the number of words on the main panel. We can see that after 100 words, the approach to $+100\%$ slows, and therefore much of the difference in average happiness between Saturdays and Tuesdays can be attributed to these first 100 words.

The third and final inset on the bottom right shows the relative total contribution of the four kinds of words to the shift in average happiness. For example, the area of the top right (yellow) circle represents the sum of all contributions due to relatively positive words that increase in frequency on Saturdays over Tuesdays ($+\uparrow$). We find that the sizes of these circles are not always transparently connected to the top 50 words, with smaller contributions combining over the full set of 1034 words. For Saturdays relative to Tuesdays, while the largest individual word contributions are due to more frequent use of positive words ($+\uparrow$), overall, the less frequent use of negative words ($-\downarrow$) has about the same impact. On the other side of the ledger, we see the smaller total contribution of words going against the trend of happier Saturdays, and the more frequent use of negative words ($-\uparrow$) slightly outweighs the drop in certain positive words ($+\downarrow$).

The two numbers above the circles give the total percentage change toward and away from the reference text's average happiness. In this example, there is a drop in happiness of -46% of $h_{\mathrm{avg}}^{(\mathrm{comp})} - h_{\mathrm{avg}}^{(\mathrm{ref})}$ due to less use of positive words, $+\downarrow$, and more use of negative words, $-\uparrow$. On the other side, more frequent positive words, $+\uparrow$, and less frequent negative words, $-\downarrow$. contribute to a rise in happiness equal to $+146\%$ of $h_{\mathrm{avg}}^{(\mathrm{comp})} - h_{\mathrm{avg}}^{(\mathrm{ref})}$. The two changes combine to give $+100\%$ of $h_{\mathrm{avg}}^{(\mathrm{comp})} - h_{\mathrm{avg}}^{(\mathrm{ref})}$.

The Simpson lexical size N_{S} (Fig. 6) shows a pattern different to that of average happiness: we observe that while Tuesday is again the overall minimum, a strong

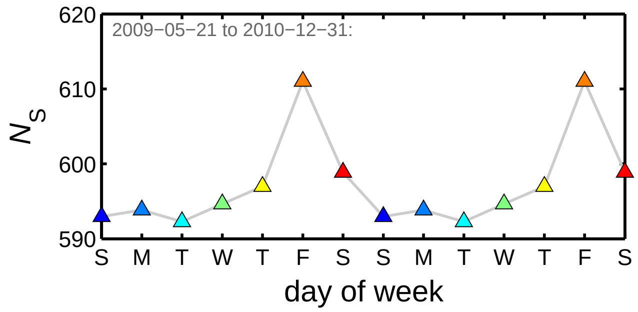

FIG. 6: Simpson lexical size as a function of day of the week.

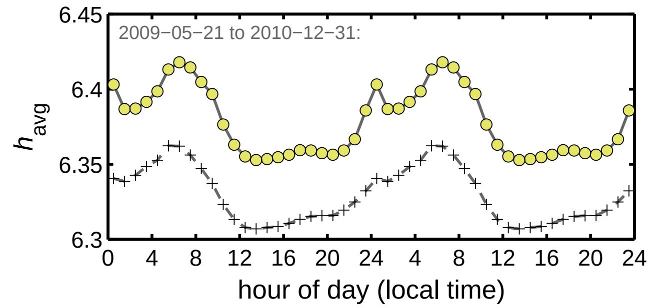

FIG. 7: Average happiness level according to hour of the day, adjusted for local time. Solid circles show h_{avg} with all data and all ANEW study words used. The curve marked with crosses shows the form of h_{avg} with the omission of two ANEW study words, 'happy' and 'birthday'. We use the same vertical scale as in Figs. 3 and 4.

maximum appears on Friday; there are mild elevations on Thursday and Saturday, and Sunday through Wednesday represents a flat low.

To see further into these changes between days, can generate word shift graphs for Simpson lexical size N_{S}. These word shift graphs are simpler than those for average happiness as they depend only on changes in word frequency. Using the definition $N_{\text{S}} = 1/S = 1/\sum_{i=1}^{N} p_i^2$, we obtain

$$N_{\text{S}}^{(\text{comp})} - N_{\text{S}}^{(\text{ref})} = \frac{1}{S^{(\text{comp})}S^{(\text{ref})}} \sum_{i=1}^{N} \left(\left[p_i^{(\text{ref})}\right]^2 - \left[p_i^{(\text{comp})}\right]^2 \right). \quad (4)$$

We next define the individual percentage contribution in the shift in Simpson lexical size as

$$\delta N_{\text{S},i} = \frac{100}{\left|S^{(\text{ref})} - S^{(\text{comp})}\right|} \left(\left[p_i^{(\text{ref})}\right]^2 - \left[p_i^{(\text{comp})}\right]^2 \right), \quad (5)$$

where $\sum_i \delta N_{\text{S},i} = \pm 100$ depending on the sign of $|S^{(\text{ref})} - S^{(\text{comp})}|$. Note that the reversal of the reference and comparison elements in Eq. (5) reflects the fact that any one word increasing in frequency decreases overall diversity. Further, no other diversity measure ($q \neq 2$) allows for a linear superposition of contributions such as we find in Eq. (5), and this was one of the reasons we provided earlier for choosing a lexical size based on Simpson's concentration.

Using Eq. (5), we find Friday's larger value of N_{S} relative to Tuesday's can be attributed to changes in the frequency of around 100 words. Most of these words are those typically found at the start of a Zipf ranking of a text, though their ordering is of interest. We see that changes in the frequency of 12 words account for 75% of the shift in N_{S}, 50 words for 96%, and 100 words for 98.5%. The words contributing the most to the shift are 'I' (32.7%), 'to' (14.4%), 'the' (6.9%), 'is' (6.9%), 'a' (6.2%), and 'you' (5.3%). The drops in 'I' and 'you' are accompanied by decreases in other personal pronouns (such as 'me', 'my', and 'your'), and, along with the other changes in word abundances, may suggest a shift in focus away from the self and toward less predictable, richer fare of Friday activities. The dominant word appearing more frequently is 'RT' (-8.4%), and there are words specific to Friday that appear more frequently than on Tuesday such as '#ff' (−4.3%), 'follow' (−1.0%), 'Friday' (−0.41%), 'weekend' (−0.25%), and 'tonight' (−0.24%) (#ff is an example of a hash tag, in this case representing a popular Friday custom of Twitter users recommending other users worth following).

VII. DAILY CYCLE

We next examine how average happiness levels change through the day, hour by hour. The curve with filled circles in Fig. 7 gives the overall trend. The isolated small peak around midnight corresponds to two ANEW study words—'happy' and 'birthday'— which appear disproportionately at the change of the day, likely due to automated, well-wishing tweets. Removing these two words gives the dashed curve with crosses, and the effect is to erase the small peak and to otherwise lower the overall happiness level in a uniform fashion across the day.

The daily pattern of happiness in tweets shows variation similar to that which we observed for the weekly cycle (Fig. 3), here ranging from a low of $h_{\text{avg}} \simeq 6.35$ between 1 and 2 pm to a high of $h_{\text{avg}} \simeq 6.42$ between 6 and 7 am. Starting from 1 pm, we see a constant level maintained through to 9 pm, when a rise in average happiness begins. The afternoon low is consistent with self-reported moods; Stone et al. in particular, observe a happiness dip in the afternoon [52]. Even with the artificial peak at midnight removed, we see an increase in happiness through the night. A small jump occurs in the change from 4–5 am to 5–6 am, which also marks 'biological midnight.' People after this point in time are more likely to be rising for the day rather than extending the previous one, and thus we have a change in the kinds of mental states represented by active users. From a high between 6 and 7 am, average happiness steadily descends to return to the afternoon trough. Interestingly, we also

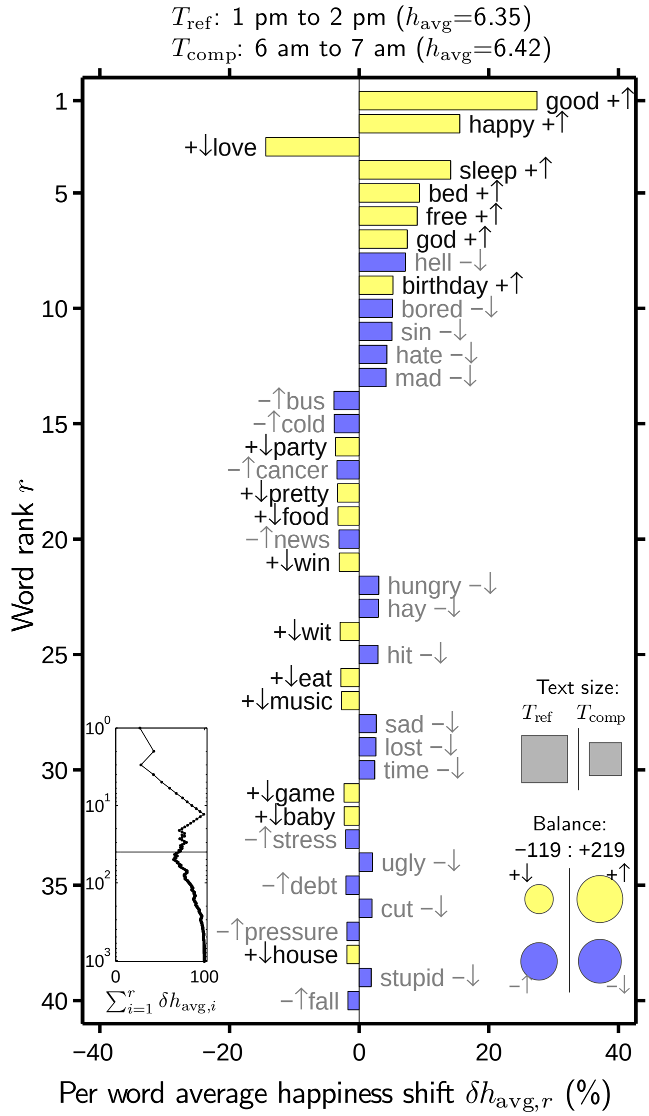

FIG. 8: Word shift graph comparing the happiest hour (6 am to 7 am) relative to the least happy hour (1 pm to 2 pm). See Fig. 5 and related text for an explanation.

find that usage rates of the most common profanities are all very similar (not shown) and are roughly anticorrelated with the observed happiness cycle. Cursing follows a sawtooth pattern with a maximum occurring around 1 am, and the lowest exhibition of swearing matching up with the daily early morning happiness peak between 6 and 7 am.

To give a deeper sense of the underlying moods reflected in the low and high of the day, we construct a word shift graph in Fig. 8, comparing tweets made in the hours of 6 to 7 am and 1 to 2 pm. As the lower left inset cumulative plot shows, the first 11 ranked positive shifts (along with one negative one, 'love') account for most of the overall shift. Thereafter, words balance back and forth with most of the difference made up by the 200th ranked word (see inset showing cumulative sum $\sum_i \delta h_{\mathrm{avg,i}}$). The

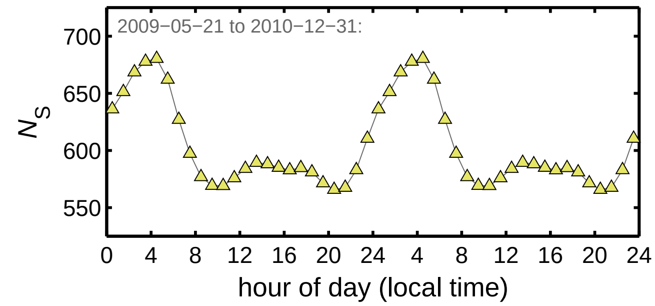

FIG. 9: Simpson lexical size N_{S} for time of day, corrected according to local time. The daily cycle is repeated twice for clarity.

first few salient, relatively positive words more abundant between 6 and 7 am (+↑) are 'good' (e.g., appearing in good morning), 'happy', 'sleep', 'bed', 'free', and 'god'. These are joined with strong decreases in negative words (−↓) including 'hell', 'bored', 'sin, and 'hate'. As shown in the figure, a few words that work against the overall difference are the more prevalent, early morning negatives such as 'bus', 'cold', and 'news' (all −↑). Also against the overall trend are positive words used less often such as 'love', 'party', and 'food (+↓).

In Fig. 9, we show that Simpson lexical size N_{S} follows a cycle roughly similar in shape to average happiness. The peak through the night is more pronounced than for happiness, taking off around 10 pm, climbing until 4 to 5 am ($N_{\mathrm{S}} \simeq 680$); from there, N_{S} drops rapidly to a local minimum in the morning (10 to 11 am), and then rises slightly to reach a minor crest in the early afternoon before slowly declining to the day's minimum between 10 and 11 pm ($N_{\mathrm{S}} \simeq 566$). In examining the change in N_{S} between the high at 4 to 5 am and the low in 10 to 11 pm, we see that just the top 20 ranked words yield 83.4% of the overall difference, and the top 100 give 96.0%. The first few contributions by rank are 'I' (30.8%), 'a' (28.3%), 'the' (16.0%), 'de' (9.9%), 'to' (7.4%), and 'que' (7.4%), which appear less frequently between 4 and 5 am. Most all other words making substantive contributions are prepositions and pronouns. The only word in the top 20 that becomes more frequent and thus effects a decrease in N_{S}, is the first ranked 'RT' (−60.9%). Users' writings thus appear to be more rich and less predictable during the night, with an apex near biological midnight. Another potential explanation may be a rise in automated tweets, perhaps purposefully introducing an element of randomness.

VIII. HAPPINESS AVERAGES AND DYNAMICS FOR KEYWORDS AND PHRASES

We turn to our last area of focus: temporal happiness patterns for tweets containing specific text elements such as words and short phrases. We need not restrict ourselves to words and indeed any text element may be considered, including, for example, dates, punctuation, emoticons, and phonemes. We examine various collections of words and text elements, ranging from long term importance ('economy'), to contemporary topics ('Obama'), to the everyday ('today'). In doing so, we are effectively generating unsolicited opinion polls regarding certain topics. Recent related work has explored correlations between public opinion polls and Twitter sentiment levels [28], as well as the use of emotional levels gleaned from Twitter to predict stock market behavior [29] with suggestive yet mixed findings. Here, we add to these findings by showing how certain happiness trends based on keywords are clearly correlated with external events. At the same time, we find many keyword-based trends are relatively stable, and our interest turns to the average happiness level which we do find to be highly variable across keywords. We begin with examples of stable time series before highlighting several volatile trends.

For keywords and phrases, we measure what we call 'ambient happiness' as follows. We first find all tweets containing a given keyword or phrase, and measure their average happiness, binning by calendar month. Further, rather than displaying absolute happiness, we create each time series as the difference in happiness between the keyword time series and the overall time series (Fig. 2), so as to remove the background trend. Note that in measuring the happiness of ambient text, we include contributions from the specific keyword in the case that we have happiness estimates for that keyword.

In Fig. 10A, we show ambient happiness time series for seven keywords and phrases, chosen so as to exhibit both a range of happiness scores and represent diverse topics and elements. The lower plot in Fig. 10B shows the relative normalized frequency of tweets containing each keyword. The trend for tweets containing the word 'happy' is to maintain a positive differential of approximately +1.3 to +1.4 above the overall average happiness time series. By contrast, the counter of 'sad' hovers just above −2.5. Words co-occurring with the emoticons ':)' and ':(' are strongly distinct in terms of happiness with means near +0.4 and −0.6, and we see a slight downward trend for ':('. Even punctuation is of interest: the exclamation mark's ambient happiness time series is a positive one though clearly below that of 'happy' and ':)'. Lastly, we show trends for two contemporary issues in the United States, 'Tea Party' and 'Afghanistan'. The phrase 'Tea Party' exhibits a fairly neutral trend, dropping gradually overall, and increasing in frequency by a factor of more than 10 starting in February, 2009. 'Afghanistan' is not surprisingly strongly negative with ambient happiness scores consistently between −1.5 and −1.

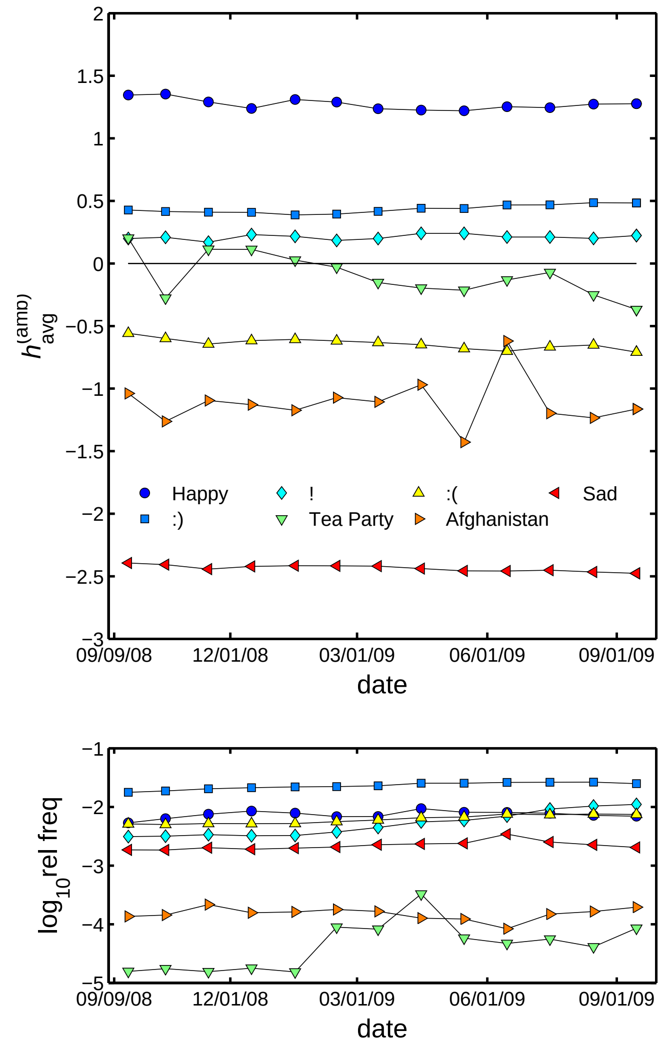

FIG. 10: **A.** Ambient happiness time series for some illustrative keywords and text elements. Ambient happiness is the average happiness of all words found occurring in tweets containing the given keyword or text element, with the background average happiness of all tweets removed (n.b., keywords are included). Binning is by calendar month and symbols are located at the center of each month. **B.** Fraction of tweets containing keywords and elements.

We have examined a selection of 100 handpicked keywords and text elements. As mentioned above, the ambient average happiness for tweets containing many of these terms are mostly stable over time, and in Tab. III we show overall ambient average happiness $h_{\rm avg}^{\rm (amb)}$ for the list, sorted in descending order. Our list is in no way exhaustive; rather it contains political keywords ('Democrat' and 'Republican'), semantic differentials ('right' and 'left'), terms relating to the economy ('money' and 'Goldman Sachs'), families of related keywords ('Jon Stewart' and 'Glenn Beck'), personal pronouns, emoticons, and so on. As such, the extremes (most and least happy words for example) are not to be presumed to remain so for larger sets of words, and our main interest is in making comparisons of related terms. In Tab. IV, we present the same terms ordered according to Simpson lexical size $N_{\rm S}$. In computing each term's $N_{\rm S}$, we exclude the term itself.

We observe many interesting patterns and the we

Word	$h_{\rm avg}^{\rm (amb)}$	Total Tweets	Total ANEW	Word	$h_{\rm avg}^{\rm (amb)}$	Total Tweets	Total ANEW
1. love	+1.42	46,687,476 (6)	85,269,499 (5)	51. me	-0.06	144,342,098 (4)	88,088,051 (4)
2. happy	+1.32	16,541,968 (13)	32,442,529 (8)	52. ?	-0.07	2,333,283 (53)	674,679 (69)
3. win	+1.26	7,981,856 (26)	14,640,728 (20)	53. commute	-0.09	90,126 (94)	90,092 (92)
4. kiss	+1.21	1,697,405 (59)	3,162,330 (48)	54. gay	-0.09	2,727,309 (47)	1,697,177 (57)
5. cash	+1.21	1,279,236 (63)	2,468,496 (51)	55. right	-0.10	19,166,480 (10)	15,850,283 (19)
6. vacation	+1.11	934,501 (67)	1,783,270 (56)	56. school	-0.11	9,264,217 (24)	6,924,193 (34)
7. Christmas	+1.03	4,887,968 (35)	10,645,630 (25)	57. Republican	-0.13	229,773 (86)	188,338 (85)
8. God	+0.95	8,576,364 (25)	17,867,768 (16)	58. they	-0.16	27,442,360 (8)	27,150,189 (11)
9. party	+0.93	6,438,886 (29)	12,090,597 (23)	59. winter	-0.19	1,255,945 (64)	1,217,225 (64)
10. sex	+0.89	3,551,767 (39)	7,087,972 (31)	60. lose	-0.19	2,056,468 (55)	2,091,540 (53)
11. Valentine	+0.85	247,288 (84)	464,914 (75)	61. Jon Stewart	-0.20	52,084 (97)	33,086 (96)
12. family	+0.79	5,014,816 (32)	10,629,361 (26)	62. gas	-0.22	1,022,879 (65)	812,029 (68)
13. sun	+0.65	2,385,348 (52)	4,602,627 (44)	63. no	-0.22	95,129,093 (5)	38,894,616 (6)
14. life	+0.50	14,006,454 (17)	27,770,768 (10)	64. Democrat	-0.23	93,193 (93)	75,450 (93)
15. hope	+0.48	11,833,337 (18)	22,952,366 (13)	65. left	-0.27	4,893,634 (34)	4,611,878 (43)
16. heaven	+0.43	741,878 (71)	1,485,702 (59)	66. Senate	-0.29	447,732 (78)	316,835 (80)
17. :)	+0.42	10,470,483 (20)	6,787,678 (35)	67. election	-0.30	560,184 (75)	375,055 (78)
18. income	+0.36	510,425 (76)	418,161 (77)	68. Sarah Palin	-0.34	225,577 (87)	150,096 (88)
19. friends	+0.33	7,669,719 (27)	7,541,106 (29)	69. Obama	-0.35	2,981,150 (44)	1,998,326 (54)
20. snow	+0.32	2,596,165 (49)	5,011,785 (40)	70. economy	-0.36	608,878 (73)	460,834 (76)
21. :-)	+0.32	1,680,165 (60)	1,102,512 (67)	71. Congress	-0.36	391,510 (79)	279,695 (81)
22. night	+0.29	17,089,505 (12)	17,606,796 (17)	72. drugs	-0.39	509,606 (77)	469,091 (74)
23. vegan	+0.28	183,889 (90)	178,676 (86)	73. Muslim	-0.42	215,300 (88)	146,506 (89)
24. Jesus	+0.27	2,027,720 (56)	1,673,992 (58)	74. George Bush	-0.43	32,341 (98)	23,102 (98)
25. girl	+0.25	10,070,132 (22)	19,886,691 (14)	75. climate	-0.44	364,177 (80)	229,129 (83)
26. USA	+0.23	2,157,172 (54)	1,204,585 (65)	76. Pope	-0.51	152,320 (91)	135,955 (90)
27. you	+0.22	173,276,993 (3)	145,464,084 (2)	77. oil	-0.53	1,377,355 (62)	1,148,990 (66)
28. our	+0.21	14,062,465 (16)	14,437,899 (21)	78. I feel	-0.54	5,173,513 (31)	4,702,352 (42)
29. ;)	+0.20	2,618,940 (48)	1,475,221 (60)	79. Glenn Beck	-0.54	113,991 (92)	101,090 (91)
30. health	+0.20	2,575,543 (50)	4,950,202 (41)	80. Islam	-0.54	187,223 (89)	70,311 (94)
31. tomorrow	+0.20	10,379,637 (21)	8,899,406 (28)	81. :-(	-0.65	341,141 (81)	244,215 (82)
32. !	+0.16	3,463,257 (40)	1,385,072 (62)	82. :(	-0.70	2,907,145 (45)	1,891,225 (55)
33. summer	+0.13	2,998,785 (43)	2,554,459 (50)	83. flu	-0.75	901,403 (68)	639,000 (70)
34. we	+0.13	39,132,934 (7)	34,513,587 (7)	84. rain	-0.78	3,233,464 (41)	5,959,903 (38)
35. today	+0.13	25,588,506 (9)	23,619,518 (12)	85. BP	-0.78	582,167 (74)	326,100 (79)
36. man	+0.12	15,856,341 (14)	29,558,118 (9)	86. mosque	-0.79	69,812 (95)	46,736 (95)
37. woman	+0.10	2,543,036 (51)	5,603,347 (39)	87. dark	-0.95	1,577,553 (61)	3,233,911 (47)
38. Stephen Colbert	+0.10	23,778 (99)	14,697 (99)	88. Lehman Brothers	-1.08	8,500 (100)	4,280 (100)
39. ;-)	+0.10	943,413 (66)	516,171 (73)	89. Goldman Sachs	-1.08	52,703 (96)	30,769 (97)
40. RT	+0.06	339,055,724 (1)	142,219,359 (3)	90. Afghanistan	-1.15	273,519 (83)	172,637 (87)
41. coffee	+0.04	2,800,972 (46)	2,399,867 (52)	91. Iraq	-1.37	238,931 (85)	213,425 (84)
42. church	+0.03	1,812,251 (58)	3,452,171 (45)	92. cold	-1.39	3,670,447 (36)	7,015,518 (32)
43. work	+0.02	18,415,618 (11)	16,191,802 (18)	93. gun	-1.81	680,903 (72)	1,263,217 (63)
44. I	+0.02	307,960,343 (2)	282,865,043 (1)	94. hate	-2.43	9,652,881 (23)	18,158,870 (15)
45. yes	+0.02	11,593,356 (19)	7,499,840 (30)	95. hell	-2.49	6,266,162 (30)	11,056,735 (24)
46. them	0.00	15,352,295 (15)	14,398,889 (22)	96. sick	-2.55	3,576,058 (37)	6,783,395 (36)
47. hot	-0.01	7,122,144 (28)	6,286,163 (37)	97. sad	-2.56	3,563,745 (38)	6,951,686 (33)
48. boy	-0.01	4,933,333 (33)	9,670,512 (27)	98. war	-2.63	1,955,901 (57)	3,417,588 (46)
49. yesterday	-0.01	3,077,761 (42)	2,852,623 (49)	99. depressed	-2.64	280,872 (82)	541,394 (72)
50. Michael Jackson	-0.02	825,979 (70)	571,442 (71)	100. headache	-2.83	856,600 (69)	1,446,064 (61)

TABLE III: A selection of 100 keywords and text elements ordered by average ambient happiness $h_{\rm avg}^{\rm (amb)}$. We define ambient happiness of a keyword as the average happiness of tweets containing that keyword, relative to the overall happiness of tweets, $h_{\rm avg} \simeq 6.37$. The number of tweets and total number of ANEW study words are listed in the third and fourth columns, with the ranking of the keyword according to these quantities shown in brackets. Note that all pattern matches with tweets were case-insensitive.

invite the reader to explore the tables beyond the observations we detail here. We begin with the highest and lowest rankings of ambient happiness $h_{\rm avg}^{\rm (amb)}$, for our list, finding them to be reassuring sensible. The three top ranked words are 'love' ($h_{\rm avg}$=+1.42), 'happy' ($h_{\rm avg}$=+1.32), and 'win' ($h_{\rm avg}$=+1.22), and the last five, in reverse order, are 'headache' ($h_{\rm avg}$=−2.55), 'depressed' ($h_{\rm avg}$=−2.56), 'war' ($h_{\rm avg}$=−2.63), 'sad'

Word	N_S	Total Words	Frac top 50K	Word	N_S	Total Words	Frac top 50K
1. RT	1019.5	4,750,853,207 (1)	0.653 (100)	51. Iraq	235.5	3,721,777 (84)	0.832 (68)
2. ?	662.1	26,075,929 (58)	0.731 (98)	52. Jon Stewart	234.9	705,345 (97)	0.836 (62)
3. !	621.1	36,823,480 (50)	0.742 (97)	53. Senate	233.7	6,790,890 (78)	0.826 (71)
4. USA	501.5	31,502,001 (54)	0.751 (94)	54. happy	232.8	204,117,404 (17)	0.834 (65)
5. no	487.3	1,431,112,579 (5)	0.763 (93)	55. climate	231.7	5,244,765 (80)	0.813 (81)
6. ;-)	476.9	13,228,710 (67)	0.75 (95)	56. yes	230.0	148,444,393 (21)	0.846 (50)
7. ;)	389.2	33,790,283 (52)	0.791 (86)	57. today	225.3	380,178,182 (9)	0.883 (20)
8. war	386.2	29,006,196 (56)	0.785 (88)	58. election	220.7	8,631,903 (75)	0.847 (47)
9. Goldman Sachs	379.5	718,341 (96)	0.766 (92)	59. summer	219.1	44,709,706 (42)	0.864 (39)
10. gay	377.6	38,233,416 (46)	0.823 (77)	60. Christmas	215.7	63,298,968 (35)	0.862 (41)
11. me	368.4	2,135,939,836 (4)	0.829 (70)	61. rain	215.1	46,201,833 (41)	0.836 (61)
12. :-)	362.3	22,795,437 (61)	0.773 (91)	62. girl	214.0	151,296,368 (20)	0.873 (32)
13. Islam	355.2	2,775,797 (89)	0.678 (99)	63. I feel	214.0	71,411,852 (34)	0.901 (4)
14. :)	347.1	131,341,152 (24)	0.775 (90)	64. kiss	212.7	24,630,584 (59)	0.845 (51)
15. Muslim	343.9	3,326,881 (86)	0.779 (89)	65. God	211.6	129,782,153 (25)	0.884 (18)
16. Michael Jackson	335.0	10,289,280 (71)	0.803 (83)	66. school	211.2	132,771,182 (23)	0.88 (25)
17. Obama	325.8	44,117,986 (43)	0.825 (74)	67. coffee	209.1	39,261,662 (45)	0.878 (27)
18. Lehman Brothers	324.5	116,106 (100)	0.743 (96)	68. Afghanistan	208.8	3,897,596 (83)	0.793 (85)
19. :-(	312.5	4,797,808 (81)	0.804 (82)	69. heaven	208.3	10,754,013 (69)	0.864 (38)
20. health	312.4	38,166,377 (47)	0.826 (72)	70. left	207.8	80,172,086 (31)	0.873 (31)
21. gas	311.8	15,802,878 (65)	0.822 (78)	71. family	207.8	77,004,415 (32)	0.873 (30)
22. Jesus	311.4	30,107,870 (55)	0.831 (69)	72. them	205.1	267,245,300 (12)	0.893 (9)
23. :(	304.5	38,023,334 (48)	0.798 (84)	73. sad	203.6	54,823,749 (36)	0.886 (17)
24. hot	298.3	98,261,855 (28)	0.847 (46)	74. night	203.1	242,877,502 (13)	0.883 (21)
25. cash	298.0	19,087,947 (63)	0.832 (66)	75. hell	202.7	90,002,715 (30)	0.883 (19)
26. vegan	290.9	2,696,126 (90)	0.845 (54)	76. mosque	198.3	1,080,516 (95)	0.82 (80)
27. George Bush	288.0	454,636 (98)	0.847 (48)	77. tomorrow	198.1	151,595,090 (19)	0.892 (11)
28. BP	285.2	8,956,756 (74)	0.791 (87)	78. friends	197.5	124,239,635 (27)	0.886 (16)
29. man	283.3	233,321,861 (15)	0.845 (52)	79. vacation	197.1	13,414,587 (66)	0.876 (28)
30. sex	276.2	51,861,876 (37)	0.844 (57)	80. snow	195.6	36,983,507 (49)	0.881 (22)
31. Sarah Palin	275.4	3,193,862 (87)	0.842 (58)	81. yesterday	192.7	50,028,287 (39)	0.887 (14)
32. we	272.4	643,395,514 (6)	0.869 (34)	82. right	190.5	285,427,508 (10)	0.887 (15)
33. flu	270.8	12,789,073 (68)	0.826 (73)	83. church	189.1	26,683,624 (57)	0.879 (26)
34. income	270.7	7,681,148 (76)	0.835 (63)	84. cold	188.4	51,159,884 (38)	0.9 (5)
35. I	269.8	4,589,834,993 (2)	0.881 (23)	85. lose	187.2	33,350,841 (53)	0.881 (24)
36. oil	267.1	21,473,375 (62)	0.825 (75)	86. sick	186.6	49,848,465 (40)	0.899 (6)
37. Democrat	262.4	1,469,218 (94)	0.832 (67)	87. economy	186.5	9,512,028 (73)	0.847 (49)
38. drugs	261.7	7,632,616 (77)	0.862 (40)	88. dark	186.1	24,032,491 (60)	0.868 (36)
39. our	257.6	239,358,233 (14)	0.869 (35)	89. Pope	185.3	2,267,900 (91)	0.84 (59)
40. boy	256.7	71,743,199 (33)	0.857 (42)	90. win	185.1	126,080,142 (26)	0.825 (76)
41. Glenn Beck	252.3	1,739,803 (92)	0.851 (44)	91. life	180.4	221,021,587 (16)	0.892 (10)
42. Stephen Colbert	251.0	297,193 (99)	0.844 (55)	92. woman	178.8	41,508,436 (44)	0.874 (29)
43. Valentine	248.4	3,168,861 (88)	0.822 (79)	93. work	178.3	279,132,525 (11)	0.898 (7)
44. party	242.9	94,658,288 (29)	0.844 (56)	94. depressed	175.2	4,108,286 (82)	0.906 (2)
45. gun	241.9	10,304,234 (70)	0.836 (60)	95. sun	166.9	36,222,335 (51)	0.849 (45)
46. winter	240.2	18,709,666 (64)	0.854 (43)	96. commute	165.0	1,469,890 (93)	0.887 (13)
47. Republican	239.8	3,607,069 (85)	0.845 (53)	97. hope	157.2	185,322,897 (18)	0.89 (12)
48. they	239.8	474,941,264 (8)	0.896 (8)	98. love	149.9	640,929,905 (7)	0.865 (37)
49. you	239.2	2,483,649,389 (3)	0.871 (33)	99. headache	126.7	10,052,304 (72)	0.907 (1)
50. Congress	236.8	6,220,609 (79)	0.834 (64)	100. hate	106.5	138,212,350 (22)	0.902 (3)

TABLE IV: The same keywords and elements as listed in Tab. III sorted according to the Simpson lexical size N_S for all tweets containing them. Keywords themselves are not included in the calculation of N_S. The third and fourth columns show the total number of words (other than the keyword) used to measure N_S and the fraction of these words that are in our fixed list of 50,000 words (the higher the better). The numbers in brackets give rankings.

(h_{avg}=−2.64), and 'sick' (h_{avg}=−2.83).

Turning to financial terms, we see tweets mentioning the dissolved firm of 'Lehmann Brothers' and 'Goldman Sachs' are both strongly negative while relatively high in lexical size ($h^{(amb)}_{avg}$=−1.08, N_S=379 and $h^{(amb)}_{avg}$=−1.08, N_S=324). With ranks of 88 and 89 in our 100 words for $h^{(amb)}_{avg}$, these companies find themselves below the average of heavy metal lyrics, 0.3 below 'BP' (British Petroleum), and above war-mired 'Afghanistan' and 'Iraq' in our table ($h^{(amb)}_{avg}$=−1.15 and −1.37). We see 'economy' is pegged at the same somewhat negative level as political terms ($h_{\text{avg}}^{(\text{amb})}$=−0.36) but conversely returns a low information level (N_{S}=186). The more personal term 'cash' appears in highly positive tweets with $h_{\text{avg}}^{(\text{amb})}$=+1.21.

Tweets referring to United States politics are below average in happiness with 'Obama', 'Sarah Palin', and 'George Bush' all registering similar scores ($h_{\text{avg}}^{(\text{amb})}$=−0.35, −0.34, and −0.43), which is on par with the New York Times (Tab. I). At the same time, these political figures all correspond to large lexical sizes (N_{S}=326, 275, and 288 respectively). A number of other political words such as 'election','senate', and 'congress' also have $-0.36 < h_{\text{avg}}^{(\text{amb})} < -0.29$. Both coming in with relatively higher average happiness, 'Republican' exceeds 'Democrat' in ambient happiness ($h_{\text{avg}}^{(\text{amb})}$=−0.13 versus −0.23) but trails in information content (N_{S}=240 versus 262).

Tweets involving the word 'war' rank high in information (N_{S}=386) and are unsurprisingly low in terms of happiness ($h_{\text{avg}}^{(\text{amb})}$=−2.63). The keywords Muslim, Islam, and mosque also register some of the lower ambient happiness scores: $h_{\text{avg}}^{(\text{amb})}$=−0.42, −0.54, and −0.79.

Generally, personal pronouns tell a positive prosocial story with 'you' and 'our' outranking 'I' and 'me' in happiness ($h_{\text{avg}}^{(\text{amb})}$=+0.22 and +0.21 versus +0.02 and −0.06). The least happy pronoun on our list is the easily demonized 'they' at −0.16, a moderate happiness level. However, tweets involving pronouns indicating self appear to be more information rich in comparison with those pointing to others: 'me' and 'we' rank 11th and 32nd (N_{S} = 368 and 272), while 'they' and 'them' rank 48th and 72nd overall (N_{S} = 240 and 205).

Tweets with 'summer' are considerably happier than those with 'winter' but less diverse: $h_{\text{avg}}^{(\text{amb})}$=+0.13 and N_{S}=219 versus $h_{\text{avg}}^{(\text{amb})}$=−0.19 and N_{S}=240. Other semantic differentials show reasonable differences. Tweets with 'hot' are happier than those with 'cold' ($h_{\text{avg}}^{(\text{amb})}$=−0.01 versus −1.39). The sequence 'yesterday', 'today', and 'tomorrow' suggests a positive view of the future with corresponding ambient happiness scores of $h_{\text{avg}}^{(\text{amb})}$=−0.01, +0.13, and +0.20.

Emoticons in increasing order of happiness are :(, :-(, ;-), ;), :-), and :) with $h_{\text{avg}}^{(\text{amb})}$ spanning −0.70 to +0.42, while in terms of information level, the order is :(, :-(, :), :-), ;), and ;-) with N_{S} ranging from 305 to 477. We see that happy emoticons correspond to higher levels of both ambient happiness and information but the ordering changes in a way that appears to reflect a richness associated with cheekiness and mischief: the two emoticons involving semi-colon winks are third and fourth in terms of happiness but first and second for information.

Tweets involving the 'fake news' comedian Stephen Colbert are both happier and of a higher information level than those concerning his senior colleague Jon Stewart ($h_{\text{avg}}^{(\text{amb})}$=+0.10 and N_{S}=251 versus $h_{\text{avg}}^{(\text{amb})}$=−0.20 and N_{S}=235). By contrast, tweets mentioning Glenn Beck are lower in happiness than both Colbert and Stewart but comparable to Colbert in information content ($h_{\text{avg}}^{(\text{amb})}$=−0.54 and N_{S}=252).

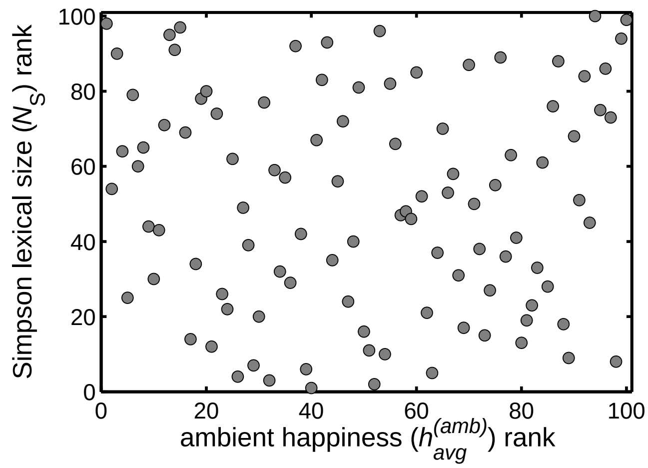

FIG. 11: For the 100 keywords and text elements listed in Tab. III, a rank-rank plot of Simpson lexical size N_{S} versus ambient happiness $h_{\text{avg}}^{(\text{amb})}$. The two variables appear to be uncorrelated, an observation supported by Spearman's rank correlation measuring $r_s = -0.016$ (p-value $\simeq 0.87$).

The exclamation point returns a positive $h_{\text{avg}}^{(\text{amb})}$=+0.16, clearly above the question mark's $h_{\text{avg}}^{(\text{amb})}$=−0.07. The gap between them is much smaller for information with both scoring highly and ranking second (?, N_{S}=662) and third (!, N_{S}=621) overall. These high values of N_{S} are sensible due to the versatility of punctuation, and RT's top ranking reflects the diverse nature of status updates shared by users.

A reflection on the preceding survey suggests that groups of related terms may possess positive, negative, or neutral correlation between happiness and information content. Overall, for our set of 100 keywords and text elements, we measure Spearman's rank correlation coefficient as $r_s = -0.016$ (p-value $\simeq 0.87$), indicating no correlation, a finding supported visually in Fig. 11. We thus have strong evidence that the two main quantities of interest that we have studied in this paper are, generally speaking, independent. Several observations follow. First of all, this independence warrants further study for other texts and, if possible, explanation. Second, both quantities (or correlated variables) should be reported in any characterization of large-scale texts. Third, for specific subfamilies of texts, any finding of a statistically and quantitatively significant correlation between happiness and lexical size is of interest and deserving of further investigation.

Finally, in Figs. 12 and 13, we present four ambient happiness time series for tweets containing the terms 'Tiger Woods', 'BP', 'Pope', and 'Israel'. For each example, we include word shift graphs that illuminate the

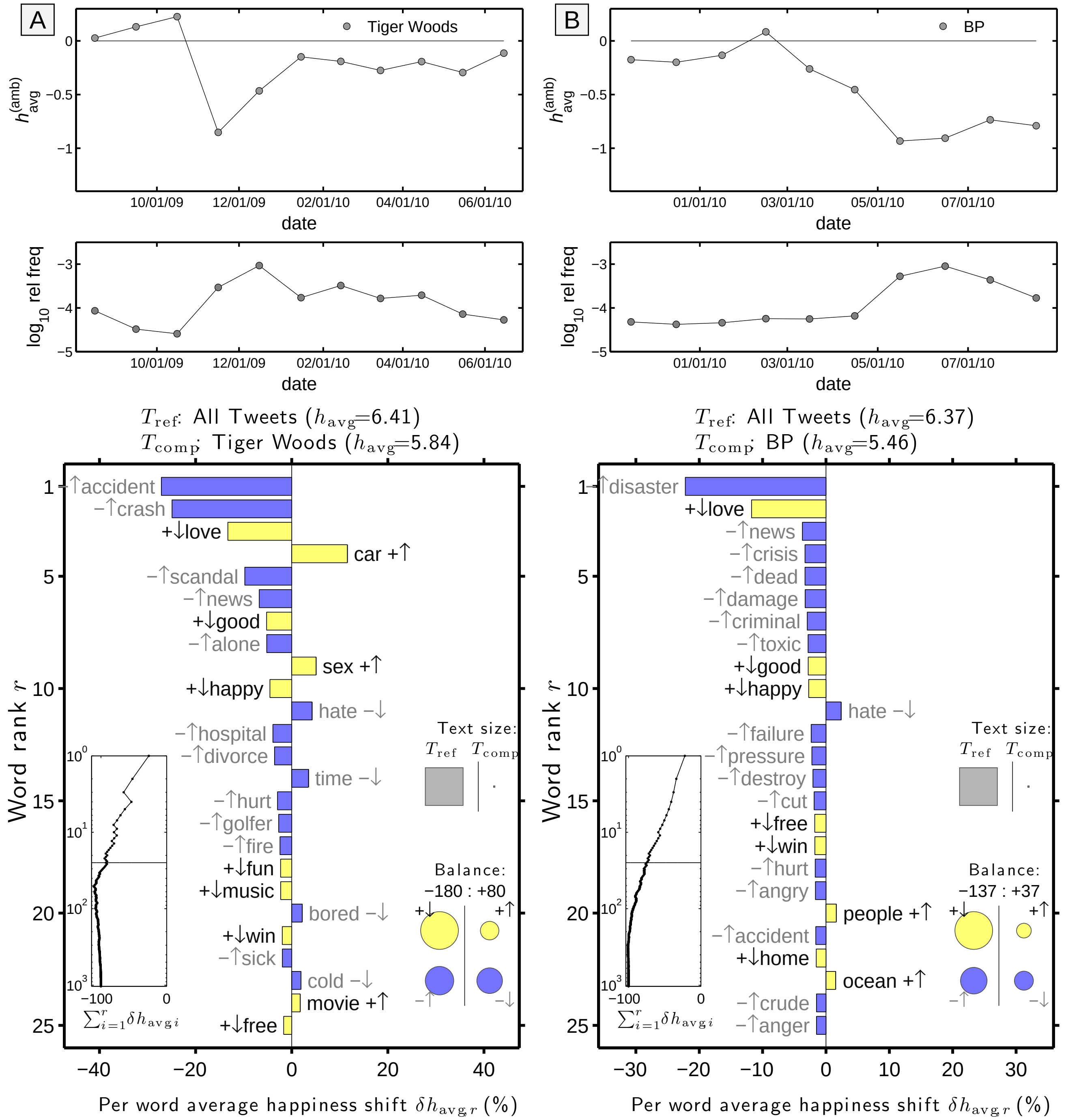

FIG. 12: Ambient happiness time series and word shift graphs for tweets containing the keywords 'Tiger Woods' and 'BP'. Ambient happiness of a keyword is $h_{\rm avg}^{\rm (amb)}$ for all words co-occurring in tweets containing that keyword, with the overall trend for all tweets subtracted. The word shift graphs are for tweets made during the worst month and the ensuing one—November and December, 2009 for 'Tiger Woods' and May and June, 2010 for 'BP.'

difference in word composition and tone for the most extreme month and the following month in comparison to that of all tweets during the same period. All of these topics involve a negative event or events leading to global media coverage.

In Fig. 12A, we show the ambient happiness time series for Tiger Woods drops abruptly in November, 2009 when his extramarital affairs famously became public after Woods crashed his car into a fire hydrant around Thanksgiving. The National Enquirer had published a claim of infidelity a few days before, and knowledge of Woods's manifold extra-marital relationships were soon

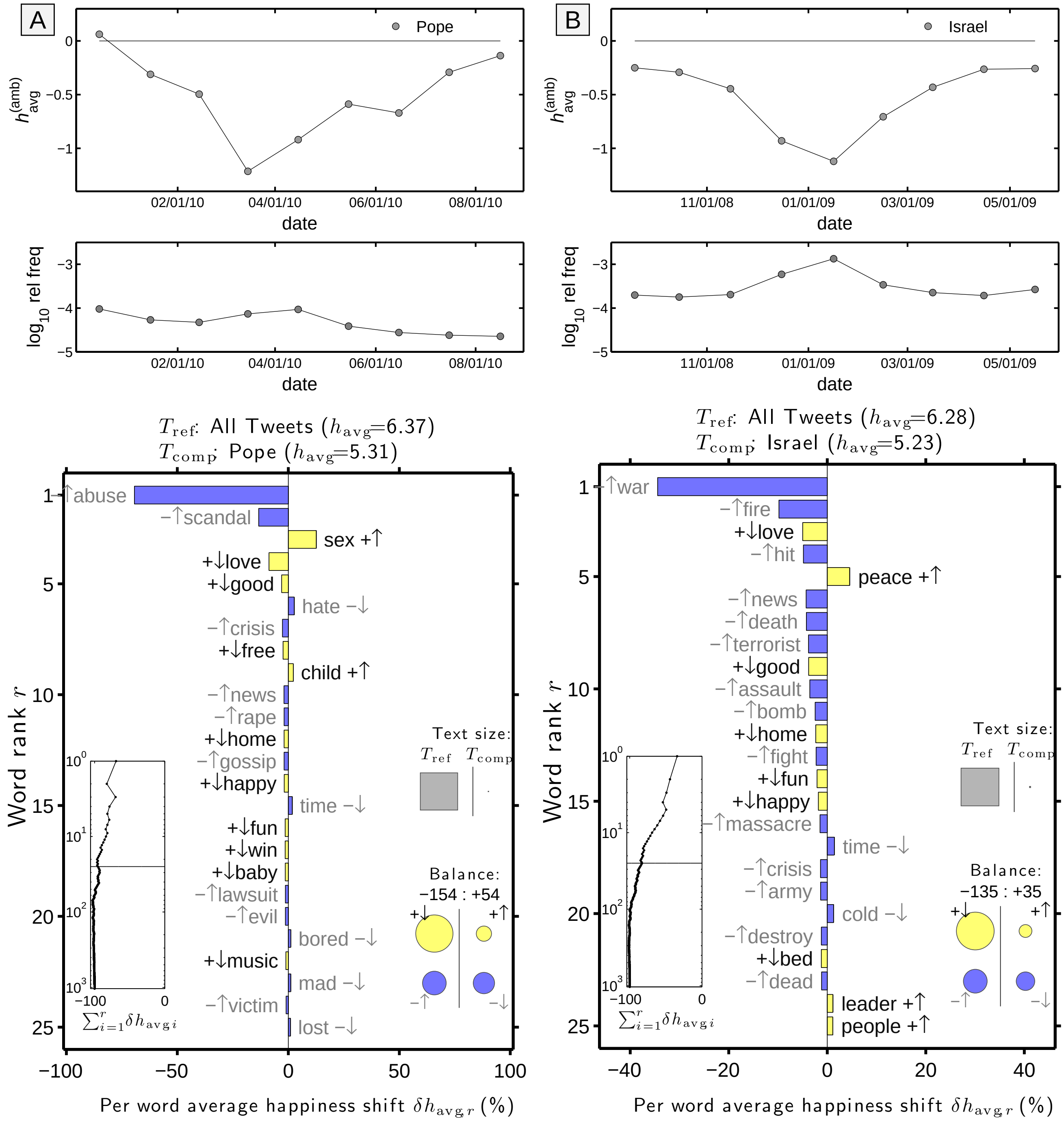

FIG. 13: Time series and word shift graphs for tweets containing the keywords 'Pope' and 'Israel'. The word shift graphs are for the time periods March and April, 2010 for 'Pope' and January and February, 2010 for 'Israel.' See Fig. 12's caption for more details.

widely being reported in the general media. Tweets concerning Woods, the world's longstanding number one golfer at the time, dropped sharply in happiness level and then rebounded over the next few months to a slightly below average steady state. The jump in media coverage is reflected in the number of tweets (middle plot). As the word shift graph shows for November and December 2009, negative words such as 'accident', 'crash', 'scandal', 'hospital', and 'divorce' pull the average happiness down below the baseline. The words 'car' and 'sex', in isolation considered to be relatively happy words, here improved h_{avg} for Woods, showing one of the potential failings of our word-centric approach. Nevertheless, the net effect is clear and such microscopic errors are overcome for large enough texts. Overall, of the four word types, the largest contribution to the drop comes from a decrease in the use of positive words ($+\downarrow$). Relatively negative words both increase ($-\uparrow$) and decrease ($-\downarrow$) in total so as to balance each other out.

In Fig. 12B, we see the decline of British Petroleum's average happiness following the April 20, 2010 explosion and collapse of the deep sea drilling platform Deepwater Horizon in the Gulf of Mexico. The well proved to be extremely difficult to cap and oil spewed into the Gulf for nearly three months. In comparing tweets containing 'BP' to all tweets in May and June, 2010, we find a drop in $h_{\rm avg}^{(\rm amb)}$ of -0.91 due to relative increases in words such as 'disaster', 'news', 'crisis', 'criminal', 'destroy', as well as 'hurt', 'angry', and 'anger', and decreases in the appearance of 'love', 'good', and 'happy'. Overall, and in contrast to the balance shown in the Tiger Woods word shift, we see the more frequent use of relatively negative words ($+\downarrow$) and the less frequent use of more positive words ($-\uparrow$) both contributed substantially to the sharp decrease in average happiness.

In Fig. 13A, we track the ambient happiness of the keyword 'Pope' over a nine month period starting with December, 2009. While the relative frequency of tweets containing 'Pope' changes little, a clear minimum in $h_{\rm avg}^{(\rm amb)}$ occurs in March, 2010. The Catholic Church's long running child molestation scandal was brought into even sharper focus during this month, notably via a Papal apology to the Irish church, and the New York Times publishing documents concerning Pope Benedict's past decisions on child molestation cases, opening up a highly charged dialogue between the media and the Vatican. In the word shift graph, we see the nadir of March and April arising from the more frequent use of negative words such as 'abuse', 'scandal', and 'crisis', and the drop in positive words such as 'love', 'good', and 'free. The increase use of the words 'sex' ($h_{\rm avg}$=8.05) and 'child' ($h_{\rm avg}$=7.08) in tweets containing 'Pope' goes against the trend (see remarks above for Tiger Woods). Although 'abuse' is the largest single contributing word ($-\uparrow$), the overall picture is similar to that for Tiger Woods: the drop in positive words ($+\downarrow$) is the main reason 'Pope' tweets are far below the average happiness level for March and April, 2010.

Our last example, Fig. 13B, shows ambient happiness for tweets involving 'Israel' from September 2008 through to May 2009. The drop in November and December reaching a minimum in January matches with the Gaza War, fought between Israel and Hamas. The increase in 'Israel' tweets also captures the increase in media reporting during this conflict. In the top ranked 25 words contributing to the strong decrease for January and February relative to the overall time series, we see the major changes primarily coming from the more frequent use of negative words ($-\uparrow$) such as 'war', 'death', 'terrorist', and 'massacre', and secondarily to drops in positive words ($+\downarrow$) 'love', 'good', and 'home'. However, when taking all ANEW study words, we see that the overall drop is due largely to a decrease in positive words ($+\downarrow$) and to a lesser extent an increase in negative words ($-\uparrow$). Last, against this rather bleak sequence of negative word shifts, we may take some solace in seeing the word 'peace' appear more often ($+\uparrow$).

IX. CONCLUDING REMARKS

In analysing temporal patterns of happiness and information content for the very large data set generated by Twitter thus far, we have been able to uncover results ranging across many timescales and topics. The weekly and daily cycles in particular appear to be robust and suggestive of universal forms, accepting that the seven day week cycle is an historical artifact.

An essential part of our comparative analyses is the word shift graph, for both happiness and information. These provide us with a detailed view of why two texts differ based on changes in word frequency. These graphs, and their future iterations, should be of use in a range of fields where size distributions are compared through summary statistics (e.g., understanding how species diversity in ecological populations may differ as a result of changes in individual species abundances).

As we have described, the metadata accompanying Twitter messages contains more information than time stamps. Future research will naturally address (and go beyond) geographic variations, particularly for the United States; the change in expressions over time for individuals and the possibility of correlation or contagion of sentiment; effects of popularity as measured by follower count on users' expressions; and the possibility of fine-scale emotional synchronization between individuals based on directed messages. Improving our simple-minded text analysis method will follow from expanding on the original ANEW study survey to encompass a much larger set of words, chosen according to frequency of use. Another step would be to accommodate common 2-grams such as 'child abuse' and 'sex scandal' as well as negated sentiments such as 'not happy'. This would improve the reliability of our happiness and information measures without losing the transparency of our current approach.

As we have seen in both the work of others and ours, Twitter and similar large-scale, online social networks have thus far provided good evidence that scientifically interesting and meaningful patterns can be extracted from these massive data sources of human behavior. The extent to which small-scale patterns can be elicited, e.g., for rare topics, also remains an open question, as does the true generalizability to the broader population. Whatever the case, Twitter is currently a substantial, growing element of the global media and is worth studying in its own right, just as a study of newspapers would seem entirely valid. And while current evidence suggests 'instant polls' created by remote-sensing text analysis methods are valid, and that these instruments complement and may in some cases improve upon traditional surveys, analysts will have to remain cognizant of the ever present problem of users gaming online expression systems to misinform.

Finally, the era of big data social sciences has undoubtedly begun. Rather than being transformed or revolutionized we feel the correct view is that the social sciences are expanding beyond a stable core to become data-abundant fields. In a data-abundant science, the challenge moves first to description and pattern finding, with explanation and experiments following. Instead of first forming hypotheses, we are forced to spend considerable time and effort simply describing. The approaches applicable for a data-scarce science still remain of the same value but new, vast windows into social and psychological behaviour are now open, and new tools are available and being developed to enable us to take in the view.

Methods

We defined a word as any contiguous set of characters bounded by white space and/or a small set of punctuation characters. Our raw word list therefore included all misspellings, words from any language used on Twitter, hyperlinks, etc. All pattern matches we made were case-insensitive, and we did not perform stemming (e.g., 'love' and 'loved' were counted separately).

The data feed from Twitter was provided in xml and json formats [46]. Early on, the data feed contained many repeated tweets, and while the fraction of duplicates dropped substantially over time, we nevertheless were obliged to check for and remove all such tweets. (Due to these various changes, all measures involving emoticons are derived from the time series up until only November 9, 2009.)

In measuring and comparing information content, a computational difficulty with the Twitter data set lies in accommodating the sheer number of distinct words. We found approximately 230 million unique words from a random sample of 25% of the tweets in our database. We determined that restricting our attention to a more manageable set of the first 50,000 most frequent words would be sufficient for highly accurate estimates of generalized entropy H_q with $q \gtrsim 1.5$, and therefore Simpson's concentration S when $q = 2$. We did not use Shannon's entropy since it converges too slowly (akin to $q = 1$) for the skew we observed in the Twitter word frequency distribution. Importantly, in fixing a list of words, we were able to account for information content differences between texts at the level of words.

Consequently, we recorded the frequencies for this specific set of 50,000 words at the level of hours and days. Note that we also always recorded the total number of words for any particular subset of tweets, so that our word probabilities were correctly normalized.

Acknowledgments

The authors thank B. O'Connor for discussions and data sharing, M. Pellon, J. Harris, and N. Riktor. The authors are grateful for the computational resources provided by the Vermont Advanced Computing Center which is supported by NASA (NNX 08A096G). PSD was supported by NSF CAREER Award # 0846668.

[1] P. Hedström, Science **311**, 786 (2006).
[2] G. Bell, T. Hey, and A. Szalay, Science **323**, 1297 (2009).
[3] A. Halevy, P. Norvig, and F. Pereira, IEEE Intelligent Systems **24**, 8 (2009).
[4] T. Hey, S. Tansley, and K. Tolle, eds., *The Fourth Paradigm: Data-Intensive Scientific Discovery* (Microsoft Research, Redmond, WA, 2009).
[5] J. P. Collins, Science Magazine **327**, 1455 (2010).
[6] Sloan Digital Sky Survey, `http://www.sdss.org/`.
[7] Large Synoptic Survey Telescope, `http://www.lsst.org/lsst`.
[8] M. Stephens, *Mapping the universe at 30 terabytes a night: Jeff kantor, on building and managing a 150 petabyte database*, The Register (2008), `http://www.theregister.co.uk/2008/10/03/lsst_jeff_kantor/`.
[9] J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, et al., Science **304**, 66 (2004).
[10] Large Hadron Collider, `http://lhc.web.cern.ch/lhc/`.
[11] `http://ngrams.googlelabs.com/`.
[12] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, T. G. B. Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, et al., Science Magazine **331**, 176 (2011).
[13] S. Goel, J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts, Proc. Natl. Acad. Sci. **107**, 17486 (2010).
[14] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, Nature **459**, 1012 (2009).
[15] H. Choi and H. Varian, Tech. Rep., Google Inc. (2009).
[16] S. Asur and B. A. Huberman, ArXiv e-prints (2010), 1003.5699.
[17] R. Layard, *Happiness* (The Penguin Press, London, 2005).
[18] D. Gilbert, *Stumbling on Happiness* (Knopf, New York, 2006).
[19] S. Lyubomirsky, *The How of Happiness* (The Penguin Press, New York, 2007).
[20] P. S. Dodds and C. M. Danforth, Journal of Happiness Studies (2009), doi:10.1007/s10902-009-9150-9.
[21] M. Bradley and P. Lang, Technical report C-1, University of Florida, Gainesville, FL (1999).
[22] F. Y. Edgeworth, *Mathematical Physics: An Essay into the Application of Mathematics to Moral Sciences* (Kegan Paul, London, UK, 1881).
[23] D. Kahneman and J. Riis, in *The science of well-being*, edited by F. A. Huppert, N. Baylis, and B. Keverne (Oxford University Press, Oxford, UK, 2005), pp. 285–304.

[24] J. H. Fowler and N. A. Christakis, BMJ **337**, article #2338 (2008).
[25] L. Jost, Oikos **113**, 363 (2006).
[26] http://www.well-beingindex.com/.
[27] http://apps.facebook.com/usa_gnh/.
[28] B. O'Connor, R. Balasubramanyan, B. Routledge, and N. Smith, in *Proceedings of 4th International AAAI Conference on Weblogs and Social Media (ICWSM-2010)* (2010).
[29] J. Bollen, A. Pepe, and H. Mao, http://arxiv.org/abs/1010.3003.
[30] J. Bollen, A. Pepe, and H. Mao, http://arxiv.org/abs/arXiv:0911.1583.
[31] M. Mitrović, G. Paltoglou, and B. Tadić, http://arxiv.org/abs/1011.6268.
[32] P. J. Rentfrow, S. D. Gosling, and J. Potter, Perspectives on Psychological Science **3**, 339 (2008).
[33] http://www.gutenberg.org/.
[34] http://www.cs.cmu.edu/~enron/.
[35] E. Sandhaus, *The New York Times Annotated Corpus*, Linguistic Data Consortium, Philadelphia (2008).
[36] C. Osgood, G. Suci, and P. Tannenbaum, *The Measurement of Meaning* (University of Illinois, Urbana, IL, 1957).
[37] L. Lee, in *Computer Science: Reflections on the Field, Reflections from the Field*, edited by C. on the Fundamentals of Computer Science: Challenges, C. S. Opportunities, and N. R. C. Telecommunications Board (The National Academies Press, 2004), pp. 111–118.
[38] Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan, Proceedings of Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005) (2005).
[39] Y. Choi, E. Breck, and C. Cardie, Conference on Empirical Methods in Natural Language Processiong (EMNLP-2006) (2006).
[40] J. A. List, Advances in Economic Analysis & Policy **6**, Article 8 (2006).
[41] L. J. Skitka and E. G. Sargis, Annu. Rev. Psychol. **57**, 529 (2006).
[42] D. Kahneman, A. B. Krueger, D. Schkade, N. Schwarz, and A. A. Stone, Science **312**, 1908 (2006).
[43] E. H. Simpson, Nature **163**, 688 (1949).
[44] C. E. Shannon, The Bell System Tech. J. **27**, 379 (1948).
[45] http://www.gigatweeter.com/counter/.
[46] http://apiwiki.twitter.com/.
[47] (2009), http://www.sysomos.com/insidetwitter/.
[48] (2009), http://sysomos.com/insidetwitter/mostactiveusers.
[49] Library of Congress will save tweets, http://www.nytimes.com/2010/04/15/technology/15twitter.html.
[50] J. Harris and S. Kamvar, *We Feel Fine: An Almanac of Human Emotion* (Scribner, New York, NY, 2009).
[51] http://www.ccs.neu.edu/home/amislove/twittermood/.
[52] A. A. Stone, J. E. Schwartz, D. Schkade, N. Schwarz, A. Krueger, and D. Kahneman, Emotion **6**, 139 (2006).