Appendix I **Data Details & Description**

In Appendix I, we include a copy of the raw POMS-Bi exam and provide selected explanation and exploratory analysis of our Twitter dataset and mood analytics dictionaries omitted from the main text.

1. Replication and Reproducibility

As Twitter's policies currently stand, we are unable to share the Twitter data used with anyone outside a select group of Princeton University and University of Tokyo academics. Furthermore, even purchasing Tweets according to the parameters we specify may not produce the same results because GNIP does not provide deleted Tweets to researchers; Tweets deleted in between our purchase date and future purchase dates will not be present for analysis.[1] To balance the privacy of subjects being researched in this study with the ultimate goal of replicable science, we embrace the philosophy of "reproducible research" laid out in Leek (2015) and make all code used to generate main text results available on GitHub.[2] Because of space considerations, we have omitted several data analyses from the final version of this document.[3] Please contact the author for access. Wherever possible we follow the standard for code and data in the social sciences as explicated by Gentzkow and Shapiro (2014). All data analysis for this essay was scripted and executed in open-source language R Version 3.3.2 "Bug in Your Hair" with GUI RStudio Version 1.0.44.

2. The POMS-Bi

We include a copy of the raw POMS-Bi exam.

[1] In fact, the reverse phenomena can occur as well: GNIP does not sell the data of Twitter users who have set their accounts to private. If enough users set their accounts to public from private between the date we purchased the Tweets and the date future researchers purchase the Tweets, then future researchers will have access to more rather than less Tweets in their dataset. Conversations with GNIP's data team have confirmed that, over time, more Tweets are typically deleted (either via direct deletion of a particular Tweet or account deletion) than are set from private to public. Therefore, we expect that future purchases of data matching the parameters specified in this article will have fewer Tweets in their dataset than we report in our dataset.

[2] It is currently available at https://github.com/shabbychef/????.

[3] In particular, we have omitted all data analyses that attempt to adjust for inflation because there were no qualitative differences between results obtained with the inflation adjusted time series and the nominal time series over 2008. All quoted prices are nominal quantities.

Figure I.2.1: This is a copy of the POMS-Bi exam; it was obtained from Marsden (1996).

3. Twitter Data Description

Because this article involves a large and heterogenous set of data, here we describe our aw data, document pre-processing steps, text data and mood time series. Therefore, we carefully describe our raw data and preprocessing steps here.

3.1 Raw Data

Our raw data for constructing explanatory variables consists of the POMS-Bi test, QDAP negation lexicon from the QDAP package, data from the R sentimentr package, Twitter data, and Google N-gram data. Unfortunately, the Twitter data, POMS-Bi test, and Google N-gram data are proprietary.

3.1.1 Twitter Data

> One of the most misleading representational techniques in our language is the use of the word "I," particularly when it is used in representing immediate experience, as in "I can see a red patch."

Philosophical Remarks

LUDWIG WIGGENSTEIN (1964)

Re-tweets, hashtags and automatic location detection features were not introduced by Twitter until 2009 and so, for this essay, we are mainly concerned with the text content and user-id of each Tweet.[4] This simplifies our data management, and these factors do not appear to correlate with the mood expressed in a Tweet. We summarize these facts in Table I.3.1 and provide a picture of a Tweet in Figure I.3.2.

Term	Description	Year Introduced
Twitter	The brand and company	2006
Tweet	An up to 140 character long text message	2006
Firehose	Constant stream of all tweets in real time	2006
Followers	Users who subscribe to tweets sent by a user	2006
Retweet	When another user relays your tweet to their followers	2009
Hashtag (#)	Identifies a topic (e.g. #earnings)	2009[12]
Mentions	Number of times others mention a specific username	2009[13]
At (@)	Identifies a username (e.g. @CNBC)	2006
Cashtag ($)	Identifies a stock ticker (e.g. $GS)	2012
Geotagging	Allows users to reveal their location (e.g. Newark, NJ)	2009

Table I.3.1: This table, modified from the table presented in Forbergskog and Rylan Blom (2013) summarizes Twitter terms and concepts circa 2013. Many of the more recent studies on the joint distribution of Twitter events and equity market activity rely on relatively recent features: for example, retweets as in Sprenger et al. (2014a) and Cashtags in Nann et al. (2013). Unfortunately, at the time TMP was conducted, many of these features had not come into existence and so we cannot easily use these in our analysis.

[4] http://qz.com/135149/the-first-ever-hashtag-reply-and-retweet-as-twitter-users-invented-them/, accessed 2015/04/06.

Figure I.3.2: This picture contains an example of a Tweet taken from http://www.educatorstechnology.com/2013/06/anatomy-of-tweet-must-see-guide-for.html. The URL was last accessed 2016/01/19. The red arrows and numbered adjoining descriptors have been added to the Tweet by Sandy Kendell.

3.1.2 Sampling Twitter Data Note

By using BMZ's filters to select Tweets, we (following BMZ) select Tweets both more likely to be written in English and to have a higher emotional content than would be obtained from a purely random sample. BMZ do not perform inference on the mood time series they construct, but this sampling scheme, a type of importance sampling, should ensure that, for any finite sample, the variance of their (and our) mood time series estimators is smaller than would be obtained from a purely random sample.[1] We purchased our data from Twitter in three sets. Our first dataset, comprising the universe of Tweets in November and December consists of 3,510,351 Tweets divided into 8,757 files, each having a ten-minute duration. Our second data set, consists of the universe of Tweets from February 14, 2008 to October 31, 2008 consists of 6,165,436 Tweets divided into 9 files, one for each calendar month. Our third dataset consists of 1,272,759 Tweets divided into 53,499 files containing 10 minutes of Tweets ranging from January 1st, 2007 to February 13th, 2008. Since our investigation is for the U.S., we recode our Tweets, originally in Greenwich Mean Time (GMT), into Eastern Standard Time (EST). Over the time frame covered in TMP, we have 8,863,296 Tweets. Since BMZ report 9,853,498 Tweets by approximately 2.7 million users, we can infer that approximately 10% of the total Tweet volume has been deleted from their dataset since 2008. Our extended sample contains 10,670,826 and consists of the universe of Tweets matching BMZ's filter less deleted Tweets from July 19th, 2007 to December 31st, 2008. Strangely, the majority of user ids appear to have been

[1] Limited resources prevented us from purchasing the unfiltered universe of Tweets, which would surely be more informative.

deleted since that time: across our entire dataset we have only 677,212 unique user ids.[2] We assume that the probability of a Tweet being deleted is independent from the sentiment content of that Tweet.[3]

3.1.3 Google Web N-Gram Data

> One of us, an undergraduate at Brown University, remembers the excitement of having access to the Brown corpus, containing one million English words. Since then, our field has seen several notable corpora that are about 100 times larger, and in 2006, Google released a trillion-word corpus with frequency counts for all sequences up to five words long. In some ways this corpus is a step backwards from the Brown Corpus: it's taken from unfiltered Web pages and thus contains incomplete sentences, spelling errors, grammatical errors, and all sorts of other errors. It's not annotated with carefully hand-corrected part-of-speech tags. But the fact that it's a million times larger than the Brown Corpus outweighs these drawbacks. A trillion-word corpus-along with other Web-derived corpora of millions, billions, or trillions of links, videos, images, tables, and user interactions - captures even very rare aspects of human behavior. So, this corpus could serve as the basis of a complete model for certain tasks - if only we knew how to extract the model from the data.
>
> *The Unreasonable Effectiveness of Data*
>
> ALON HALEVY, PETER NORVIG AND FERNANDO PEREIRA ON THE GOOGLE N-GRAM CORPUS (2009)

To construct our mood index, we begin with the composed-anxious lexicon of POMS-Bi, consisting of six composed terms and six anxious terms, used by BMZ. We follow BMZ by augmenting the POMS-Bi lexicon using Google N-grams data. In January 2006, Google "took a picture" of the public web and extracted the text of all publicly available English-language webpages. It provided this data in the form of the Google N-gram corpus and temporarily made it publicly available.[4] The corpus consists of 1 trillion words organized into 1-grams, 2-grams, 3-grams, 4-grams and 5- grams. Following BMZ, we only use the 4-grams and 5-grams to augment our lexicon. We have a total of 1,313,818,354 4-grams and 1,176,470,663 5-grams for a total of 2,490,289,017 4 and 5-grams. The Google N-gram corpus is a good choice for augmenting the POMS-Bi because it has been used by Google in thousands of other NLP projects, and BMZ are chiefly interested in how words from the POMS-Bi lexicon are used in natural language.[5] Because we are uncertain of how exactly the GPOMS uses co-occurrences between the POMS-Bi terms

[2] It is well-known that, almost from the inception of Twitter, many of the most prolific Tweeters have been spambots. We strongly suspect that many of the users in BMZ's original dataset were robots that that have since been deleted by Twitter but have no means to investigate this hypothesis. Circa 2008 when BMZ collected their Tweets, Twitter was fairly ineffective at systematically detecting and deleting spambots. http://www.dailymail.co.uk/sciencetech/article-2722677/Rise-Twitter-bots-Social-network-admits-23-MILLION-users-tweet-automatically-without-human-input.html and http://www.newyorker.com/tech/elements/the-rise-of-twitter-bots, accessed 2015/01/19.

[3] In fact, as noted in Carmona (2014) tests of our linear models in Section 4 require only that the event of deletion be uncorrelated with sentiment. Nonparametric models, like those estimated in §4.3, will generally (but not always) require the stronger, independence assumption. We make the stronger assumption throughout for simplicity and convenience. Cursory investigation suggested that too few Tweets had been deleted since we purchased our dataset from Twitter to conduct a full survival analysis of deleted Tweets that might justify these assumptions.

[4] The full explanation Google N-gram dataset, samples and a link to where the corpus can be purchased may be found here: http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.htm, accessed last 2014/05/30.

[5] http://micarum.blogspot.com/2012/02/rethinking-of-sentiment-analysis.html, accessed last on 2014/11/13.

and phrases in the Google N-gram corpus in the summer of 2014 three Princeton University undergraduates created a composed-anxious lexicon by manually analyzing the Google N-gram corpus.

3.1.3 Making Mood Dictionaries

> Like everything metaphysical, the harmony between thought and reality is to be found in the grammar of the language.

Philosophical Investigations

LUDWIG WITTGENSTEIN (1953)

As mentioned above, our chief innovations over previous replication attempts is that we use the universe, rather than a subsample, of data matching BMZ's specification and we choose a time frame that includes TMP's sample as a subset. To attempt to overcome our uncertainty about exactly how GPOMS generates a "calm"-ness time series, we apply a number of functional forms to the word counts obtained from counting the number of matches between words in our dictionaries and each Tweet. We then select only the time series matching stylized visual features of the CALM time series presented in TMP for analysis. Because previous studies have identified adjectives as containing more information than other terms and we want to minimize the noise in our simple models (Wang et al., 2012) each dictionary is comprised only of adjectives, adjective phrases, and past-participles.

To generate an extended POMS-Bi composed-anxious lexicon, we isolated every 4-gram and 5-gram database in the Google N-gram database containing one of the composed-anxious POMS-Bi terms using term matching. This yielded 227,037 4-grams and 205,945 5-grams that contain a "composed" term and 109,668 4-grams and 103,449 5-grams that contain an "anxious" term respectively. Then, three Princeton University undergraduate students inspected each of the 646,099 total 4 and 5 grams. Each undergraduate assigned adjectives, adjective phrases and past-participles that they identified as having a composed and anxious connotation to a new list of composed and anxious terms, respectively. The final composed lexicon, termed ℓ_I^C, is an 81 term list obtained from the union of each of the new composed term lists, as well as the original six terms from POMS-Bi composed half-scale. The final anxious lexicon, termed ℓ_I^A, is an 78 term list obtained from the union of each of the three new anxious term lists, as well as the original six terms from POMS-Bi anxious half-scale. Like BMZ, we extend our POMS-Bi lexicon using only co-occurrences between Google N-grams database terms with terms already in the POMS-Bi composed-anxious lexicon. Our word lists are available in the public repository for this project.

3.1.4 Negation and Valence Shifting

> Bart: Are you licking toads?
> Homer: I'm not NOT licking toads.

Missionary Impossible

THE SIMPSONS (2000)

We want to be able to correctly score sentences like "I am not excited". If our system ignores the possibility of negation, it will score this sentence as anxious when in reality the author is composed. Furthermore, Bollen claims in public talks that the GPOMS can account for negation and so we want to include this feature in our scoring scheme.[6] We use the QDAP package's negation words after removing all punctuation as our negation word list when applying word count methods. The *sentiment_by* function we use to generate a secondary mood time series in the main text automatically takes negation into account.

3.1.5 Tweet Data Cleaning and Preprocessing

For Tweet cleaning for our raw word count scoring, we follow the procedure of Bollen et al. (2011a), removing all punctuation, lowercasing words, removing stop words[7] and standardizing spacing between words. All words are stemmed using the Porter (1980) stemmer, which removes affixes according to a fixed set of rules. The Porter stemmer was chosen because it is the most popular stemmer in sentiment analysis applications and because BMZ, as far as we know, rely exclusively on this stemming tool in their other, better-documented mood analysis projects. After stemming, some of the unigrams in our lexicon do not necessarily have a unique representation. For instance, "unexcited" and "unexcitable" both become "unexcit" after stemming. By stemming the words in our dataset we reduce its size, which makes computation over the dataset easier. Better yet, we reduce its dimensionality since many terms, unstemmed, can carry the same exact sentiment e.g. "happily", "happy" and "happiness" will all be assigned to our stemmer to the term "happi". All three terms likely signify "Happy" and so we can see that stemming reduces the complexity of our statistical model in a way that conforms to our intuitive notion of what a sentiment or mood is.[8] A more subtle point is that the statistical power of our tests increase under stemming

[6] https://www.youtube.com/watch?v=n0it1M0vILs, accessed 2015/01/12.
[7] Stop words are simply common words like "the", "a" and "I" that are unlikely to contain information when our documents are rendered as a BOW. We use the top 100 words used in English reading and writing as documented by Fry (1957). To construct our stop word list, we simply take this 100 word list and remove all negation words from it: in this case only "no" and "not". This gives us a 98 word list of stop words.
[8] In our negation lexicon, only one word, "nobody", is affected by the stemming process. The Porter stemmer maps "nobody" to "nobodi."

not only because we have reduced the dimensionality of our model but because we have increased the number of observations for each parameter.

4. Twitter as a Survey Tool

> ...there are at least two good reasons to suspect that this result may not be all it seems. The first is the lack of plausible mechanism: how could the Twitter mood measured by the calmness index actually affect the Dow Jones Industrial Average up to six days later? Nobody knows.
>
> ---
>
> MIT TECHNOLOGY REVIEW on TMP

Bollen and Mao (2011b) describe TMP's methodology as analogous to a poll in which they use individuals' Tweets as though they were self-reported answers to survey questions about those individuals' mood. Given this analogy, it is strange that BMZ provide little information about the individuals from whom the GPOMS estimates mood. In this section, we characterize the sample of Tweets and users from our study.

4.1 How Many Moody Tweets Are There?

Across our entire sample, we have 2,136,868 Tweets containing at least one term from word lists constructed over the course of this project. While our median and modal users in our sample do not Tweet any mood terms, our data is not sparse. In the time period from which BMZ sample 342,255 Tweets, we find 1,366,000 Tweets containing terms from our composed-anxious lexicons, respectively. Since our lexicons collectively contain less than a third of the terms in BMZ's full GPOMS lexicon and less than four percent of the terms their OF lexicon, we conclude it is very unlikely that BMZ used the universe of Tweets matching expressions in the GPOMS and OF lexicons for their Granger causality analysis. Throughout our sample, the number of Tweets per day, the number of users in our sample Tweeting and the Tweets containing terms from our lexicon are increasing, but oscillate with regularity. We capture these stylized facts in Figure I.4.I.

To better understand what goes into our time series variables, we visualize our Tweet text. Optimal text visualization is a current topic of research, but one of the most popular ways to visualize text is through word clouds (O'Connor, 2014). In Figures I.4.1 to I.4.3, we plot word clouds made from our sample.

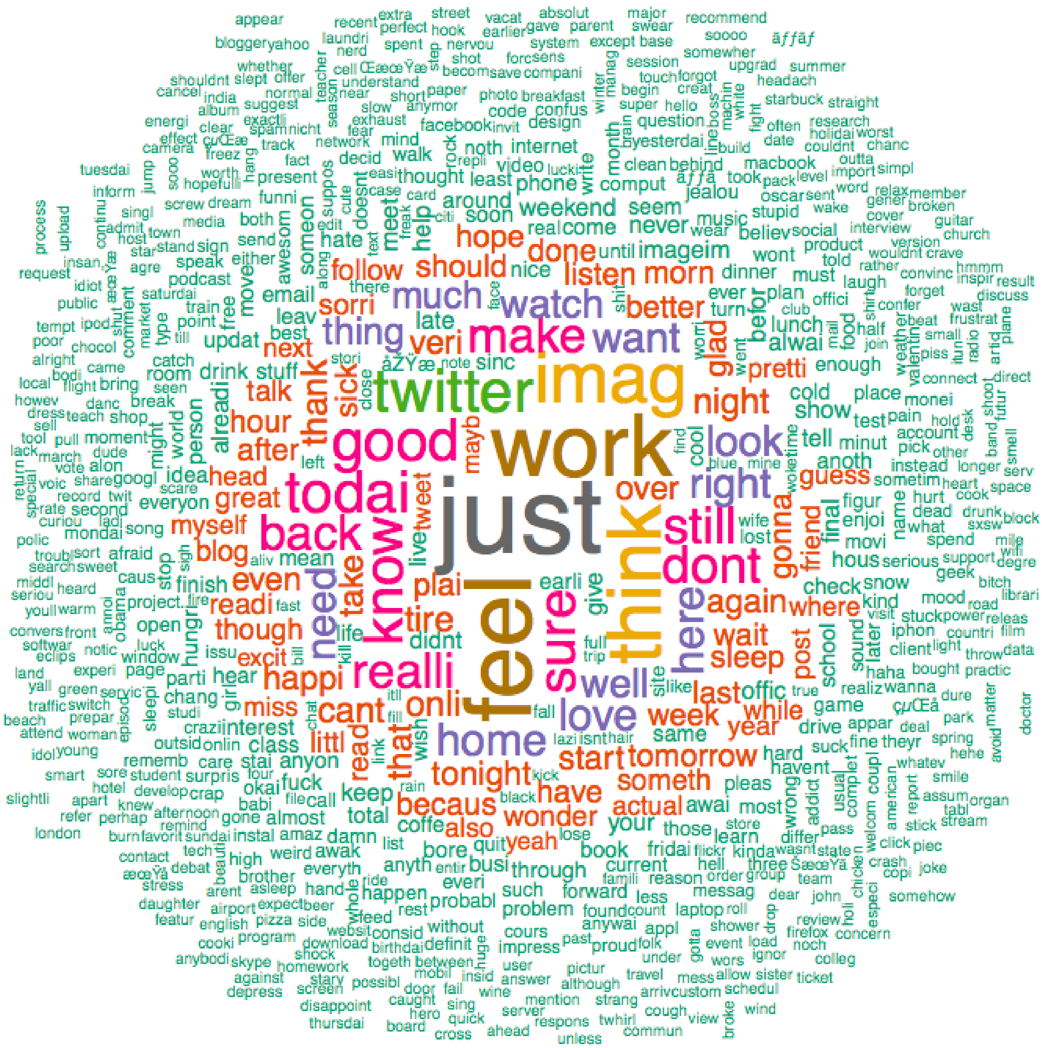

Figure I.4.1: This word cloud is made from all the Tweets we have in February, 2008. Our word clouds are color coded so that words with frequencies of about the same order of magnitude share the same color. Word sizes are scaled by word count so that the most common words shown are approximately eight times larger than the least common words shown. The effects of stemming on the dataset are clearly shown: "imag" and "todai" are in our word cloud instead of the words "image" and "today". The words closest to the center of this and all following word clouds are the most common words in our cleaned dataset with more than three characters. This word cloud contains approximately 300 tokens. Words occurring less than five times in February are not plotted.

Figure I.4.2: Clockwise from the top left panel are word clouds from a random sample of 137,049 Tweets from March, April, June and May. Our word clouds are color coded so that words with frequencies of about the same order of magnitude share the same color. Word sizes are scaled by word count so that the most common words shown are approximately six times larger than the least common words shown. Words occurring less than five times in a given month are not plotted. Notice that we can now observe that our word distributions appear approximately stationary among the most common words in our sample.

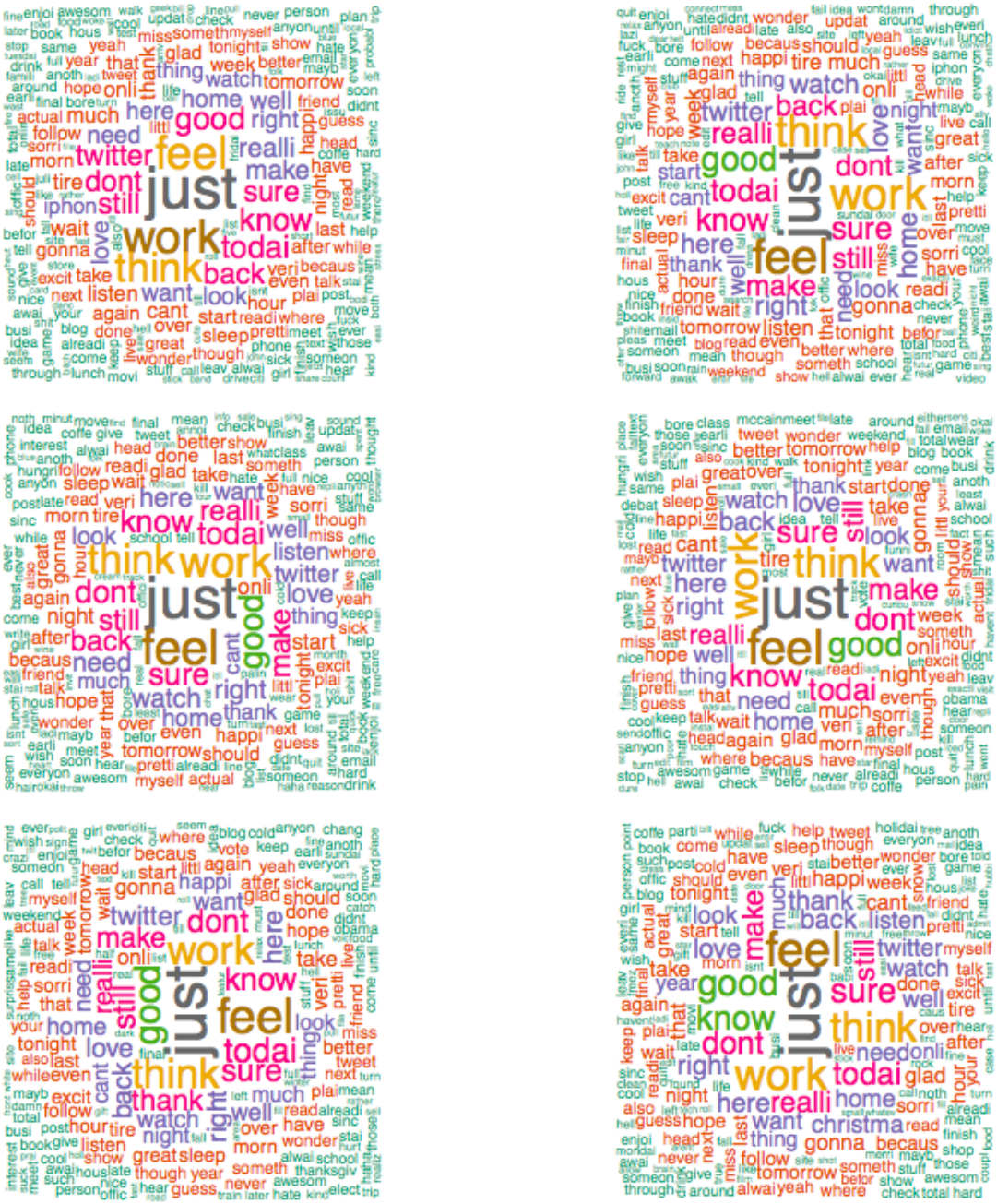

Figure I.4.3: The six word clouds above are, from left to right and then top to bottom, made from the text of a random sample of 137,049 Tweets from July, August, September, October, November and December. Our word clouds are color coded so that words occurring in our sample with frequencies of about the same order of magnitude are given the same color. The words are scaled by word count so that the most common words shown are approximately six times larger than the least common words shown. Words occurring less than five times in a given month are not plotted. These six word clouds provide more evidence of the stationarity of word frequencies.

BMZ's choice to weight by Tweet rather than user does not appear to significantly affect the daily value of the collective mood index.[1] In a sample of 100,000 users who had used at least one term from our word lists across our dataset, the median, modal and mean number of Tweets containing mood terms in our mood sample on a given day was 1, 1 and 1.02, respectively. The maximum number of Tweets containing mood terms contributed by one user on a given day in our sample is three. Unfortunately, users can share a username and one user can have many usernames. This means that, if we estimate our daily mood score from a random sample of usernames rather than Tweets, it is unclear how this mood score would relate to that obtained from a survey taken over those same individual users. By taking a sample over Tweets, BMZ obtain results approximating what they would obtain from a sample of usernames. Because they lack access to the function mapping usernames to individual identities, one cannot be certain that their approach actually estimates an average collective mood state.

4.2 Where Are These People?

An objection to TMP raised numerous times by journalists was that many of the Twitter users in TMP's sample were from outside of the United States and so it is implausible that their collective mood could impact U.S. stock markets. BMZ acknowledge this criticism, writing:

> ...our analysis is not designed to be limited to any particular geographical location nor subset of the world's population. This approach may be appropriate since the US stock markets are affected by individuals worldwide, but for the particular period under observation Twitter.com users were *de facto* predominantly English speaking and located in the US.

BMZ provide no evidence in TMP for these claims. Our word clouds show that nearly all of the text in our data is in English, however the Figures in that section do not confirm whether or not the Tweets themselves came from the United States. It is plausible that the text in these word clouds might have come from early adopters of Twitter in New Zealand, Australia, England or Canada, rather than the United States. Through correspondence with GNIP salespeople we have confirmed that in 2008, the majority of Tweets were written in English by users in the United States. From mid-2009 onward, a substantial and increasing proportion of Tweets, even those written in English, were produced by users outside the U.S. Using Tweets from this time onward, without considering geographic information, is unlikely to produce a reasonable collective mood index for the U.S.[2]

[1] However, we wonder how much of this is due to Twitter's increasing ability to identify and delete robot accounts. BMZ hint that their mood measurement system, at the time of its deployment, was vulnerable to being driven up and down by "astroturfing" and "Twitter bombing" campaigns. Given the large number of deleted users in the dataset, and the possibility that these deleted users were automated spam accounts, we are more concerned about the sensitivity of BMZ's mood index to attack by botnets.

[2] This is particularly problematic for the attempted ex-sample replications after TMP was published.

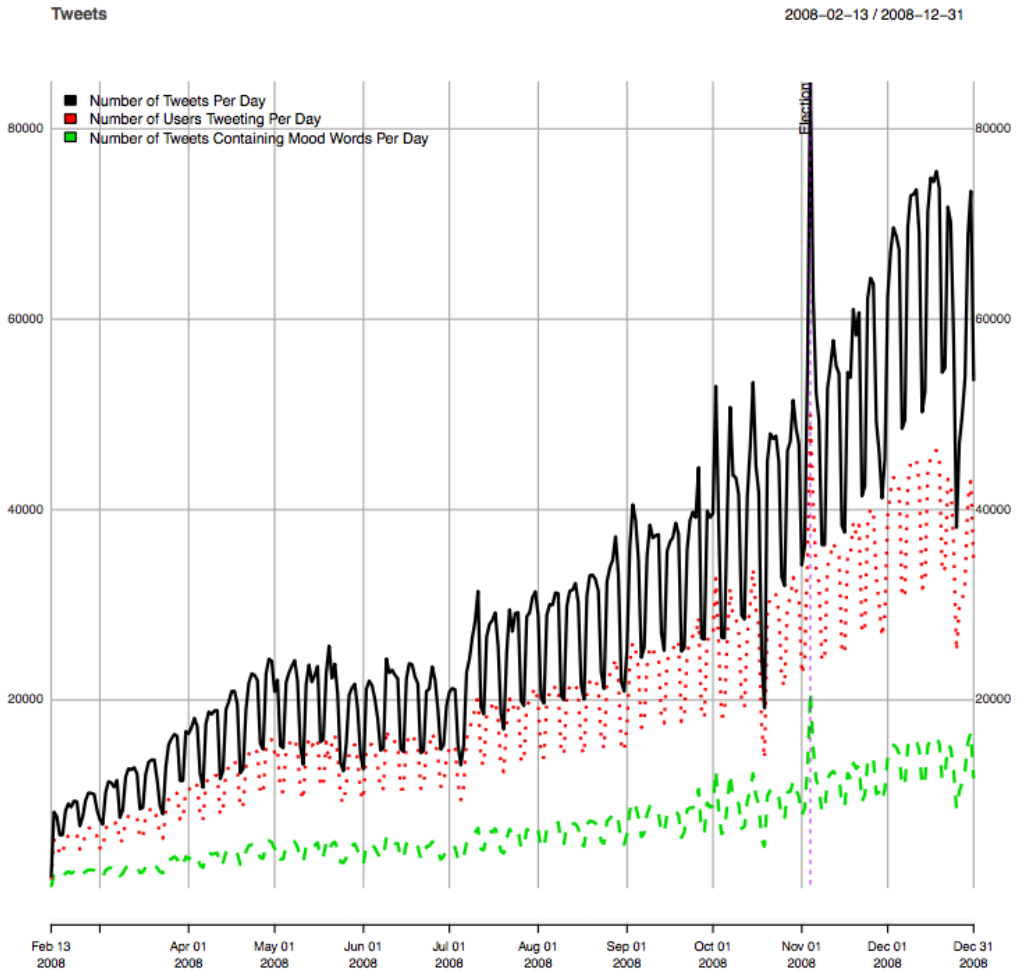

Figure I.4.4: This figure shows the number of Tweets per day on the black, solid line and the number of users Tweeting per day on the red, dotted line. The number of Tweets containing terms from our composed-anxious lexicon; the number of Tweets per day contributing to our mood statistics is on the dotted green line. The 2008 Presidential Election, 11/04/2008 is a global maximum for all three time series and is marked with a thin, vertical, purple dotted line.

However, even if this is the case for the universe of Tweets, how can one know if BMZ's sample has these characteristics? The importance sampling scheme used in TMP seems likely to generate samples de facto dominated by the U.S. citizens. Even though Twitter had not yet implemented geotagging, users in 2008 were able to self-report their location. We take two random samples of 2500 users from December, 2008. For the first sample, we plot each user's self-reported location in Figure I.4.5.[1] Since the foreign user base was growing faster than the domestic user base throughout our sample time period, the proportion of users in our sample from the United States in December, 2008 is a plausible lower bound for the proportion of users across the whole sample.[2] One may wonder if this visualization is representative of the users contributing to our collective mood measure. To answer this question, we drew another sample of 2500 Twitter users circa December, 2008, this time limiting ourselves to only those users who used one or more terms in our composed-anxious lexicon that month. Table I.4.1 presents the exact count of self-reported locations from that sample.

These results provide additional evidence for BMZ's contention that their sample consists mostly of individuals from the United States. Furthermore, Table I.4.1 suggests that the bulk of the variation in the collective mood index we derive is from individuals Tweeting in the U.S. Some commentators have objected to TMP because the sample of Twitter users is not representative of the U.S. population, based on characteristics including age, income and sociological factors. Many of the users of Twitter in 2008 were digital natives and early adopters: they are likely to display a much younger and wealthier profile than the average individual across the United States as a whole. As in Lachanski (2014) we do not find these criticisms warranted for econometric and philosophical reasons. Econometrically, if the aforementioned sources of mood error are unbiased (i.e. have mean zero for each day), then our estimated mood coefficients are biased downward. With more accurate mood measurements we are likely to find a larger DJIA effect, not a smaller one. Philosophically, it may simply be the case that Twitter users' collective mood (i.e. Twitter mood), rather than the collective mood of the nation, predicts the DJIA. Digital natives might be more likely to be marginal, price-determining investors, for instance. Following BMZ, we do not attempt to correct for the skewed demographic of Twitter.

[1] We took a sample of 2500 rather than our entire dataset because at the time the data was analyzed Google's mapping API limited free users to a sample of 2500 per day.

[2] Because the locations are self-reported, many of them are spelled wrong or use incorrect acronyms. We attempt to overcome this through the application of regular expressions.

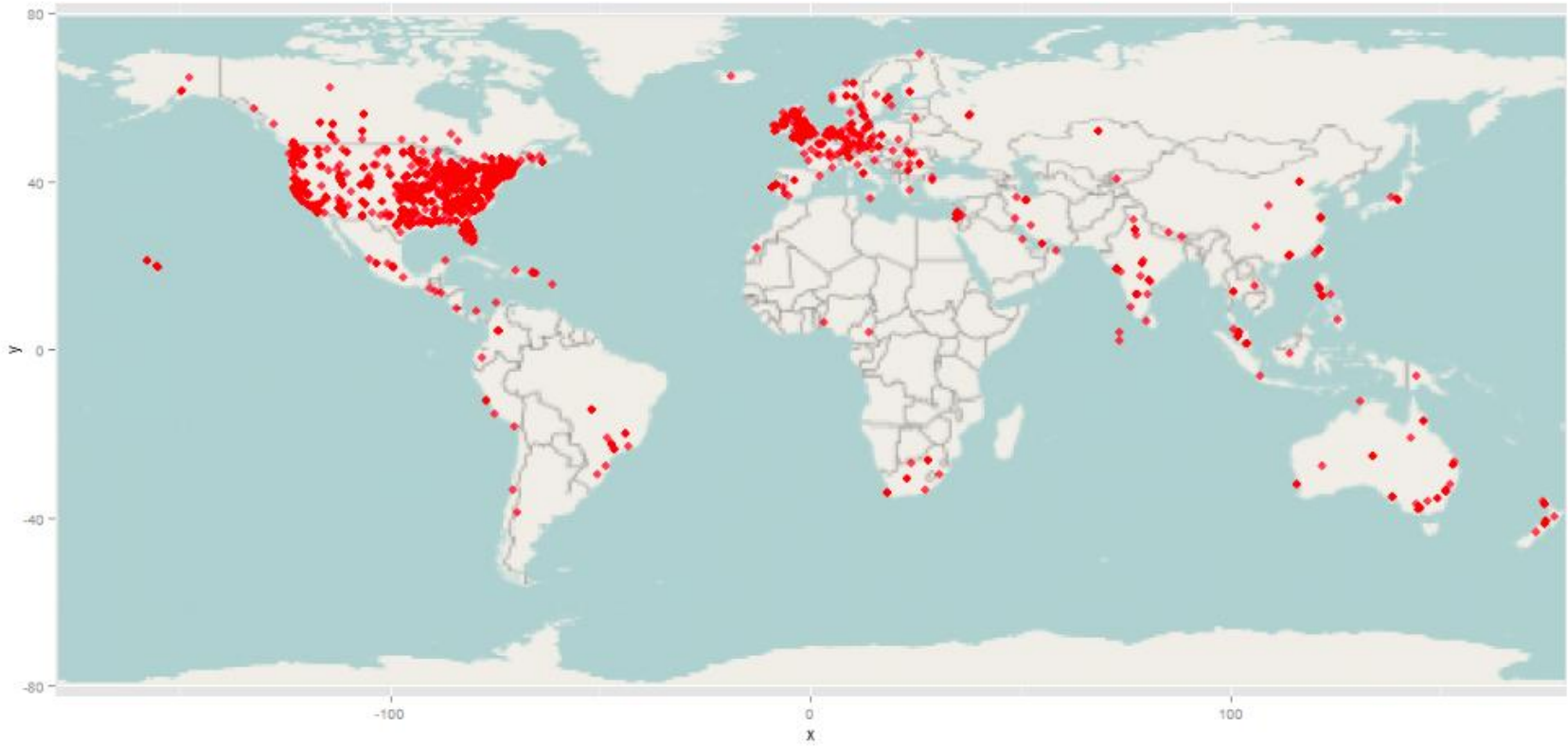

Figure I.4.5: This map visualizes the self-reported locations of 2500 users on Twitter circa December, 2008. Each red diamond represents the longitude and latitude of one user in our sample; because the map uses longitude and latitude coordinates to plot locations, several red dots completely overlap. Notice that it is dominated by the United States: 1998 users in this sample of 2500 are from the United States.

Country	# Users	Country	# Users	Country	# Users
United States of America	1734	Japan	5	Dominica	1
United Kingdom	155	New Zealand	5	Dominican Republic	1
Canada	113	Colombia	4	Ecuador	1
Australia	55	Malaysia	4	Greece	1
Germany	48	Spain	4	Guatemala	1
India	25	Czech Republic	3	Hungary	1
France	18	Hong Kong S.A.R.	3	Isle of Man	1
South Africa	15	Kazakhstan	3	Jersey	1
Netherlands	12	Portugal	3	Jordan	1
Brazil	11	Puerto Rico	3	Laos	1
Norway	11	Russia	3	Lithuania	1
China	10	Sweden	3	Maldives	1
Mexico	10	Thailand	3	Malta	1
Philippines	10	Bulgaria	2	Nepal	1
Singapore	10	Chile	2	Nigeria	1
Ireland	9	Denmark	2	Oman	1
Belgium	8	El Salvador	2	Panama	1
Italy	8	Finland	2	Poland	1
Romania	8	Indonesia	2	Republic of Serbia	1
Switzerland	7	Peru	2	Sri Lanka	1
Israel	6	Taiwan	2	Turkey	1
Austria	5	Argentina	1	United Arab Emirates	1
Iran	5	Cameroon	1	Western Sahara	1
Costa Rica	1	Croatia	1	**Unknown**	**118**

Table I.4.1: This table contains a sample of 2500 Twitter users, drawn from those who used one or more terms in ℓ_1, with self-reported locations circa December, 2008, ordered by number of people reporting from that location, then alphabetized. Because the number of non-US users grew faster than US users throughout this time period, this sample contains a plausible estimate of the upper bound of non-US users in the whole sample. This suggests that, de facto, the vast majority of users in our collective mood sample, at minimum 70%, are from the United States.