On the Study of Diurnal Urban Routines on Twitter

Mor Naaman
SC&I
Rutgers University
mor@rutgers.edu

Amy X Zhang
Computer Laboratory
University of Cambridge
amy.xian.zhang@gmail.com

Samuel Brody
Google*
sdbrody@gmail.com

Gilad Lotan
Social Flow, Inc.
gilad@socialflow.com

Abstract

Social media activity in different geographic regions can expose a varied set of temporal patterns. We study and characterize diurnal patterns in social media data for different urban areas, with the goal of providing context and framing for reasoning about such patterns at different scales. Using one of the largest datasets to date of Twitter content associated with different locations, we examine within-day variability and across-day variability of diurnal keyword patterns for different locations. We show that only a few cities currently provide the magnitude of content needed to support such across-day variability analysis for more than a few keywords. Nevertheless, within-day diurnal variability can help in comparing activities and finding similarities between cities.

Introduction

Social media activity in different geographic regions expose a varied set of temporal patterns. In particular, Social Awareness Streams (SAS) (Naaman, Boase, and Lai 2010), available from social media services such as Facebook, Twitter, FourSquare, Flickr, and others, allow users to post streams of lightweight content artifacts, from short status messages to links, pictures, and videos, in a highly connected social environment. The vast amounts of SAS data reflect, in new ways, people's attitudes, attention, and interests, offering unique opportunities to understand and draw insights about social trends and habits.

In this paper, we focus on characterizing social media patterns in different urban areas (US cities), with the goal of providing a framework for reasoning about activities and diurnal patterns in different cities. Using Twitter as a typical SAS, previous research studied *specific* temporal patterns that are *similar* across geographies, in particular in respect to expression of mood (Golder and Macy 2011; Dodds et al. 2011). We aim to provide insights for reasoning about diurnal patterns in different geographic (urban) areas that can be used in studying activity patterns in these areas, going beyond previous work that had mostly examined *topical* differences between posts in different geographic areas (Eisenstein et al. 2010; Hecht et al. 2011) or briefly examined broad diurnal differences (Cheng et al. 2011) in volume between cities. Such study can contribute to urban studies, with implications for diverse social challenges such as public health, emergency response, community safety, transportation, and resource planning as well as Internet advertising, providing insights and information that cannot readily be extracted from other sources.

Developing such a framework presents a number of challenges, both technical and practical. First, SAS data (and in particular Twitter) has been shown to be quite noisy. Users of SAS post different type of content, from information and link sharing, to personal updates, to social interactions, and many others (Naaman, Boase, and Lai 2010). Can stable patterns be reliably extracted given this noisy environment? Second, reliably extracting the location associated with Twitter content is still an open problem, as we discuss below. Finally, Twitter content volume shifts over time as more users join the service, and fluctuates widely in response to breaking events and other happenings, from Valentine's Day to the news about Bin Laden's capture and demise. Such temporal volume fluctuations might distort otherwise stable patterns and make them difficult to extract.

In this paper, therefore, we report on a study that extracts and reasons about stable temporal patterns from Twitter data. In particular, we: 1) use large scale data with manual coding to get a wide sample of tweets for different cities; 2) study within-day and across-day variability of patterns in cities; and 3) reason about differences between cities with respect to overall patterns as well as individual ones.

Related Work

Broadly speaking, this work is informed by two key areas of related work: the use of new technologies and data sources for urban studies, and studies of social media to extract "real world" insights, or temporal dynamics. Here we broadly address these areas, before discussing other recent research that directly informed our work.

The related research area sometimes dubbed "urban sensing" (Cuff, Hansen, and Kang 2008) analyzes various new datasets to understand the dynamics and patterns of urban activity. Most prominently, mobile phone data, mainly proprietary data from wireless carriers (e.g., calls made and positioning data) help expose travel patterns and broad spatio-temporal dynamics, e.g., in (Gonzalez, Hidalgo, and Barabasi 2008). Social media was also used to augment

*Amy and Sam were at Rutgers at the time of this work.

our understanding of urban spaces. Researchers have used geotagged photographs, for example, to gain insight about tourist activities in cities (Ahern et al. 2007; Crandall et al. 2009; Girardin et al. 2008). Twitter data can augment and improve on these research efforts, and allow for new insights about communities and urban environments. More recently, as mentioned above, researchers had examined differences in social media content between geographies using keyword and topic models only (Eisenstein et al. 2010; Hecht et al. 2011). Cheng et al. (2011) examine patterns of "checkins" on Foursquare and briefly report on differences between cities regarding their global diurnal patterns.

Beyond geographies and urban spaces, several research efforts have examined social media temporal patterns and dynamics. Researchers have examined daily and weekly temporal patterns on Facebook (Golder, Wilkinson, and Huberman 2007), and to some degree Twitter (Java et al. 2007), but did not address the stability of patterns, or the differences between geographic regions. Recently, Golder and Macy (2011) have examined temporal variation related to Twitter posts reflecting mood, across different locations, and showed that diurnal (as well as seasonal) mood patterns are robust and consistent in many cultures. The activity and volume measures we use here are similar to Golder and Macys, but we study patterns more broadly (in terms of keywords) and in a more focused geography (city-scale instead of timezone and country).

Identifying and reasoning about repeating patterns in time series has, of course, long been a topic of study in many domains. Most closely related to our domain, Shimshoni et al. (Shimshoni, Efron, and Matias 2009) examined the predictability of search query patterns using day-level data. In their work, the authors model seasonal and overall trend components to predict search query traffic for different keywords. The "predicitability" criteria, though, is arbitrary, and only used to compare between different categories of use.

Definitions

We begin with a simple definition of the Twitter content as used in this work. **Users** are marked $u \in U$, where u can be minimally modeled by a user ID. However, the Twitter system features additional information about the user, most notably their hometown location ℓ_u and a short "profile" description. **Content items** (i.e., tweets) are marked $m \in M$, where m can be minimally modeled by a tuple (u_m, c_m, t_m, ℓ_m) containing the identity of the user posting the message $u_m \in U$, the content of the message c_m, the posting time t_m and, increasingly, the location of the user at the time of posting ℓ_m. This simple model captures the essence of the activity in many different SAS platforms, although we focus on Twitter in this work.

Using these building blocks, we now formalize some aggregate concepts that will be used in this paper. In particular, we are interested in content for a given geographic area, and examining the content based on diurnal (hourly) patterns. We use the following formulations:

- $M_{G,w}(d, h)$ are content items associated with a word w posted in geographic area G during hour h of day d.
- $X_{G,w}(d, h) = |M_{G,w}(d, h)|$ defines a time series of the volume of messages associated with keyword w in location G.

In other words, $X_{G,w}(d, h)$ would be the volume of messages (tweets) in region G that include the word w and were posted during hour $h = 0 \dots 23$ of day $d = 1 \dots N$, where the N is the number of the days in our dataset. We describe below how these variables are computed.

Data Collection

In this section, we describe the data collection and aggregation methodology. We first show how we collected a set of tweets M_G for every location G in our dataset. We then describe how we select a set of keywords w for the analysis, and compute the $X_{G,w}(d, h)$ time series data for each keyword and location.

The data was extracted from Twitter using the Twitter Firehose access level, which, according to Twitter, "Returns all public statuses." Our initial dataset included all public tweets posted from May 2010 through May 2011.

Tweets from Geographic Regions

To reason about patterns in various geographic regions, we need a robust dataset of tweets from each region G to create the sets $M_{G,w}$. Twitter offers two possible sources of location data. First, a subset of tweets have associated geographic coordinates (i.e., geocoded, containing an ℓ_m fields as described above). Second, tweets may have the location available from the profile description of the Twitter user that posted the tweet (ℓ_u, as described above). The ℓ_m location often represents the location of the user when posting the message (e.g., attached to tweets posted from a GPS-enabled phone), but was not used in our study since only a small and biased portion of the Firehose dataset (about 0.6%) includes ℓ_m geographic coordinates. The rest of the paper uses location information derived from the user profile, as described next. Note that the profile location is not likely to be updated as the users move around space: a "Honolulu" user will appear to be tweeting from their hometown even when they are (perhaps temporarily) in Boston. Overall, though, the profile location will reflect tendencies albeit with moderate amount of noise.

We used location data associated with the profile of the user posting the tweet, ℓ_u, to create our datasets for each location G. A significant challenge in using this data is that the location field on Twitter is a free text field, resulting in data that may not even describe a geographic location, or describe one in an obscure, ambiguous, or unspecific manner (Hecht et al. 2011). However, according to Hecht et al., 1) About $\frac{1}{9}$ of users had an automatically-updated profile location (updated by mobile Twitter apps such as Ubertweet for Blackberry), and 2) 66% of the remaining users had at least *some* valid location information in their profile; about 70% of those had information that exceeded the city-level data required for this study. In total, the study suggests that 57% of users would have some profile location information appropriate for this study. As Hecht et al. (2011) report, this user-provided data field may still be hard to automatically and robustly associate with a real-world location. We will overcome or avoid some of these issues using our method specified below.[1]

To create our dataset, we had to resolve the free-text user profile location ℓ_u associated with a tweet to one of the cities (geographic areas G) in our study. Our solution had the following desired outcomes, in order of importance:

- high precision: as little as possible content from outside that location.
- high recall: as much as possible content without compromising precision.

In order to match the user profile location to a specific city in our dataset, we used a dictionary-based approach. In this approach:

1. We used a different Twitter dataset to generate a comprehensive dictionary of profile location strings that match each location in our dataset. For example, strings such as "New York, NY" or "NYC" can be included in the New York City dictionary.
2. We match the user profile associated with the tweets in our dataset against the dictionary. If a profile location ℓ_u matches one of the dictionary strings for a city, the tweet is associated with that city in our data.

We show next how we created a robust and extensive location dictionary for each city.

Generating Location Dictionaries The goal of the dictionary generation process was to establish a list of strings for each city that would reliably (i.e., accurately and comprehensively) represent that location, resulting in the most content possible, with as little noise as possible.

To this end, we obtained an initial dataset of tweets that are likely to have been generated in one of our locations of interest. We used a previously collected dataset of Twitter messages for the cities in our study. This dataset included, for every city, represented via a point-radius geographic region G, tweets associated with that region. The data was collected from September 2009 until February 2011 using a point-radius query to the Twitter Search API. For such queries, the Twitter Search API returns a set of of geotagged tweets, and tweets geocoded by Twitter to that area (proprietary combination of user profile, IP, and other signals), and that match that point-radius query. For every area G representing a city, we thus obtained a set L_G of tweets posted in G, or geocoded by Twitter to G. Notice that other sources of approximate location data (e.g., a strict dataset of geocoded Twitter content from a location) can be alternatively used in the process described below.

From the dataset L_G, we extracted the profile location ℓ_u for the user posting each tweet. For each city, we sorted the various distinct ℓ_u strings by decreasing popularity (the number of unique tweets). For example, the top ℓ_u strings for the point-radius area representing New York City included the strings *New York* (14,482,339 tweets), *New York, NY* (6,955,481), *NYC* (6,681,874) and so forth, including also subregions of New York City such as *Harlem* (587,430).

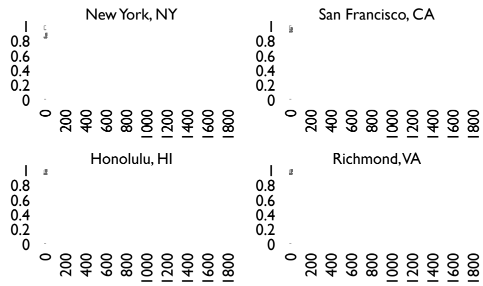


Figure 1: CDFs of location field values for the most popular 2000 location strings for four cities

We proceeded to clean the top location strings lists for each city from noisy, ambiguous, and erroneous entries. Because the distribution of ℓ_u values for each location was heavy-tailed, we chose to only look the top location strings that comprised 90% of unique tweets from each city. Figure 1 shows the cumulative distribution function (CDF) for the location string frequencies for four locations in our dataset. The x-Axis represents the location strings, ordered by popularity. The y-Axis is the cumulative portion of tweets the top strings account for. For example, the 500 most popular location terms in New York account for about 82% of all tweets collected for New York in the L_G dataset. Then, for each city, we manually filtered the lists for each city, marking entries that are ambiguous (e.g., "Uptown" or "Chinatown" that Twitter coded as New York City) or incorrect, due to geocoder issues, e.g. "Pearl of the Orient" that the Twitter geocoder associates with Boston, for one reason or another; see Hecht and Chi's (2011) discussion for more on this topic. This annotation process, performed by one of the authors, involved lookups to external sources if needed to determine which terms were slang phrases or nicknames for a city as opposed to errors.

For this study, we selected a set of US state capitals, adding a number of large US metropolitans that are not state capitals (such as New York and Los Angeles), as well as London, England. This selection allowed us to study cities of various scales and patterns of Twitter activity.

As mentioned above, using these lists of ℓ_u location strings for the cities, we filtered the Twitter Firehose dataset to extract tweets from users whose location fields *exactly* matched a string on one of our cities' lists. Using this process, we generated datasets of tweets from each city, summarized in Figure 2. The figure shows, for each city, the number of tweets in our 380-day dataset (in millions). For example, we have over 150 million tweets from Los Angeles, an average of about 400,000 a day. Notice the sharp drop in the number of tweets between the large metropolitans to local centers (Lansing, Michigan with 2.5 million tweets, an average of 6,500 per day). We only kept locations with over 2.5 million tweets, i.e. the 29 locations shown in Figure 2.

[1] In an alternative approach, a user's home location can be estimated to some degree from the topics they post about on their account (Cheng, Caverlee, and Lee 2010; Eisenstein et al. 2010; Hecht et al. 2011). For example, Cheng et al. (2010) claim 51% within-100-miles accuracy for users in their study.

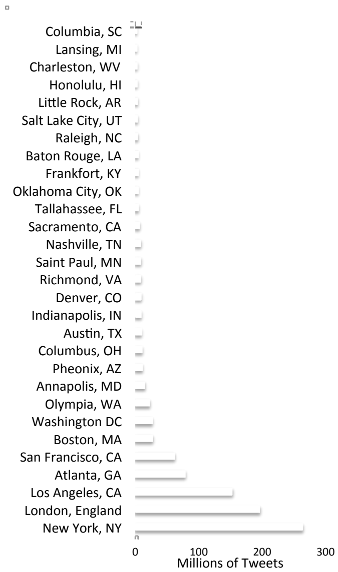


Figure 2: Location and number of associated tweets in our dataset, based on user location field matching.

Figure 3: Sparklines for volume by hour for the cities in our dataset (midnight to 11pm)

Keywords and Hourly Volume Time Series

We limited our analysis in this work to the 1,000 most frequent words appearing in our dataset (i.e., in the union of sets M_G for all locations G). We then calculated the frequency for all words in the dataset, removing posts with user profile language field that was not English. We also removed stopwords, using the NLTK toolkit. All punctuation and symbols, except for "@" and "#" at the beginning of the term and underscores anywhere in the term, and all URLs, were removed. We removed terms that were not in English, keeping English slang and vernacular intact.

In each geographic region, for each of the 1,000 terms, we created the time series $X_{G,w}(d,h)$ as described above, capturing the raw hourly volume of tweets during the days available in our dataset. We converted the UTC timestamps of tweets to local time for each city, accounting for daylight savings. For each city, we computed the time series representing the total hourly volume for every geographic region, denoted as $\mathcal{X}_G(d,h)$. Data collection issues resulted in a few missing days in our data collection period.

For simplicity of presentation, we only use weekdays in our analysis below as weekend days are expected to show highly divergent diurnal patterns. We had a total of 272 weekdays in our dataset for each city.

Diurnal Patterns

The goal of this analysis is to examine and reason about the diurnal patterns of different keywords, for each geographic region. We first examine the overall volume patterns in each location, partially to validate the correctness of the collection method for our location dataset.

Natural Periodicity in Cities

As expected, the overall volume patterns of Twitter activity for the cities in our dataset show common and strong diurnal shifts. Figure 3 shows sparklines for the average hourly proportion of volume, for all cities in our data, averaged over all weekdays in our dataset. In other words, the figure shows the average over all days of $\frac{\mathcal{X}_G(d,h)}{\sum_{i=0}^{23}\mathcal{X}_G(d,i)}$ for $h = 0, \ldots, 23$ (midnight to 11pm), where $\mathcal{X}_G(d,h)$ is the total volume of tweets in hour h of day d. The minimum and maximum points of each sparkline are marked. The figure shows the natural rhythm of Twitter activity, which is mostly similar across the locations in our dataset, but not without differences between locations. As would be expected, in each location, Twitter experiences greater volume during the daytime and less late at night and in early morning. The lowest level of activity, in most cities, is at 4-5am. Similarly, most cities show an afternoon peak and late night peak, where the highest level of activity varies between cities with most peaking at 10-11pm. However, the figure also demonstrates differences between the cities.

The consistency of patterns in the different cities is evidence that our data extraction is fairly solid, where most tweets included were produced in the same timezone (at the very least) as the city in question. Another takeaway from Figure 3 is that Twitter volume alone is probably not a immediate indication for the urban "rhythm" of a city: in other words, it is unlikely that the rise in late-night activity at cities like Lansing, Michigan is due to Lansing's many nightclubs and bars (there aren't many), but rather to heightened virtual activity that may not reflect the city's urban trends. We turn, therefore, to examining specific activities in cities based on analysis of specific keywords.

Keyword Periodicity in Cities

We aim to identify, for each city, keywords that show robust and consistent daily variation. These keywords and their daily temporal patterns represent the pulse of the city, and can help reason about the differences between cities using Twitter data. These criteria should apply to the identified keywords in each location:

- The keyword shows daily differences of significant relative magnitude (peak hours show significantly more activity).

- The keyword's daily patterns are consistent (low day-to-day variation).
- The keyword patterns cannot be explained by Twitter's overall volume patterns for the location.

Indeed, we were only interested in the keyword and location pairs that display periodicity in excess of the natural rhythm that arises from people tweeting less during sleeping hours. This daily fluctuation in Twitter volume means that daily periodicity is likely to exist in the series $X_{G,w}(d,h)$ for most keywords w and cities G. To remove this underlying periodicity and normalize the volumes, we look at the following time series transformation for each G, w time series:

$$g_{G,w}(d,h) = \frac{X_{G,w}(d,h)}{\mathcal{X}_G(d,h)} \tag{1}$$

The g transformation captures the keyword w hourly volume's portion of the total hourly volume during the same hour d, h. Here, we also account for the drift in Twitter volume over the time of the data collection. Similar normalization was also used in (Golder and Macy 2011; Dodds et al. 2011). From here on, we use Equation 1 in our analysis. In other words, we will refer to the time series $\widehat{X}_{G,w}(d,h)$ which represents the g transformation applied to the original time series of volume values, $X_{G,w}(d,h)$.

Based on the $\widehat{X}$ time series we further define the diurnal patterns series $X24_{G,w}(h)$ as the normalized volume, in each city, for every keyword, for every one hour span (e.g. between 5 and 6pm), across all weekdays in the entire span of data. In other words, for a given G, w:

$$X24_{G,w}(h) = \frac{\sum_{d=1}^{n} \widehat{X}_{G,w}(d,h)}{n} \text{ for } h = 0 \ldots 23 \tag{2}$$

In the following, we use $X24$ as a model in an information-theoretic sense: the series represents expected values for the variables for each keyword and location. We then use $X24$ and the complete $\widehat{X}$ series to reason about the information embedded within days, and across days, in different locations and for different keywords. These measures give us the significance and stability (or variability) of the patterns, respectively.

Variability Within Days

To capture variations from the model across all the days in our data we used the information-theoretic measure of entropy. Using the $X24(h)$ values, we calculate $-\sum_{h=0}^{23} p(h)log(p(h))$ where for each h, $p(h) = \frac{X24_{G,w}(h)}{\sum_{i=0}^{23} X24_{G,w}(i)}$ (the proportion of mean relative volume for that hour, to the total mean relative volumes). Entropy captures the notion of how much information or structure is contained in a distribution. Flat distributions, which are close to uniform, contain little information, and have high entropy. Peaked distributions, on the other hand, have a specific structure, and are non-entropic. We treat X24 as a distribution over relative volume for each hour to get entropy scores for each keyword.

We have experimented with alternative measures for within-day variance. For example, Fourier transforms, which can decompose a signal from its time domain to its frequency domain, was not an affective measure for this study: almost all keywords show a significant power around the 24 hour Fourier coefficient, even when using the $\widehat{X}$ series that controls for overall Twitter volume.

Variability Across Days

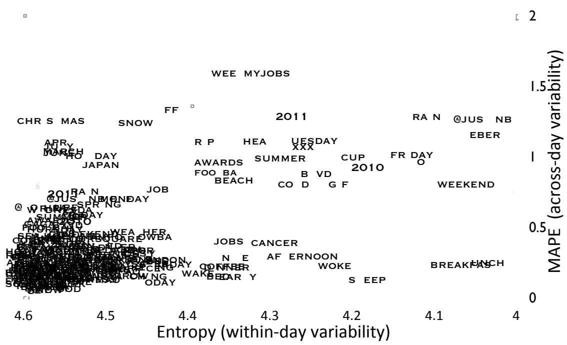
To capture variability from the model across all the days in our data we used Mean Absolute Percentage Error (MAPE) from X24, following Shimshoni et al. (2009). For a single day in the data, MAPE is defined as: $MAPE_{G,w}(d) = \sum_{i=0\ldots23} \frac{|X24(i) - \widehat{X}(d,i)|}{X24(i)}$. We then use $MAPE_{G,w}$, the average over all days of the MAPE values for the keyword w in location G.

We have experimented with alternative measures for across-day variability, including the Kullback-Leibler (KL) Divergence and the coefficient of variation (a normalized version of the standard deviation), time series autocorrelation, and the average day-to-day variation. For lack of space, we do not report on these additional measures here but note that they can provide alternatives for analysis of across-day variation of diurnal patterns.

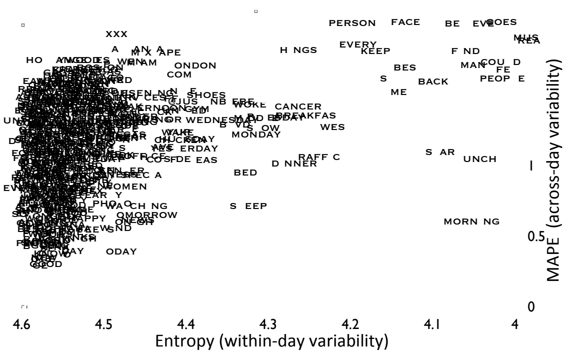
Analysis

We used these measures to study the diurnal patterns of the top keywords dataset described above for the 29 locations in our dataset. In this section we initially address a number of questions regarding the use and interpretation of the patterns: What is the connection between within- and across-day variability, and what keywords demonstrate each? How significantly is the variability influenced by volume, and is there enough volume to reason about diurnal patterns in different locations? How do these patterns capture physical-world versus virtual activities, and can these patterns be used to find similarities between locations?

Figure 4 shows scatter plots of keywords w for tweets from locations G=New York, in 4(a) and G=Honolulu, in 4(b). The words are plotted according to their within-day variability (entropy) on the x-Axis, where variation is higher for keywords to the right, and across-day variability (MAPE) on the y-Axis, where keywords with higher variability are higher. Pointing out a few examples, the keyword "lunch" shows high within-day variability and relatively low across-day variability in both New York and Honolulu: the keyword demonstrates stable and significant daily patterns. On the other hand, the keyword @justinbieber is mentioned with little variation throughout the day, but with very high across-day variability. Both parts of Figure 4 show a sample area magnified: in 4(a), an area with high across-day variability and low within-day variability, and in 4(b), an area with both low across-day variability and low within-day variability. A number of key observations can be made from Figure 4. First, there are only a few keywords with significant within-day variability, and most of them demonstrate low variation across days. The Spam keyword #jobs shows higher than usual inter-day variability amongst the keywords with high within-day variability. Second, perhaps as expected, the highly stable keywords are usually popular, generic keywords that carry little discriminatory information. Third, patterns in the data can be detected even for a low-volume location such as Honolulu (with 1% of the tweet volume of New York in our dataset). Notice, though, that the set of keywords with high within-day variation (left) in the Honolulu scatter are much more noisy, and that the across-day variation in Honolulu is higher for all keywords (the axes scales in figures 4(a) and 4(b) are identical).



(a) New York, NY



(b) Honolulu, HI

Figure 4: Within-day patterns (entropy) versus across-day patterns (MAPE) for keywords in New York City and Honolulu

How much information is needed in one area to be able to robustly reason about stability of patterns? Given the differences in MAPE values shown between the cities in Figure 4, we briefly examine the effect of volume on the stability of patterns.

Figure 5 plots the cities in our dataset according to the volume of content for each location (on the x-axis) and the number of keywords with MAPE scores below a given threshold (y-Axis). For example, in Los Angeles, with 150 million tweets, 955 out of the top 1000 keywords have MAPE values lower than 50% (i.e. relatively robust), and 270 of these keywords have MAPE lower than 20%. These values leave some hope for detection of patterns that are stable and reliable in Los Angeles – i.e., the MAPE outliers. However,

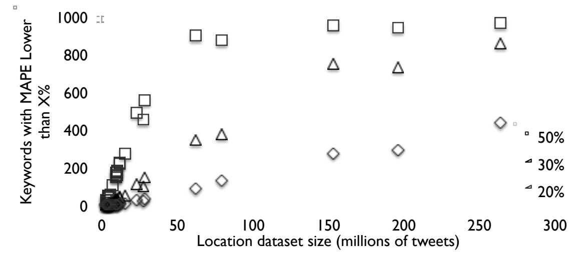


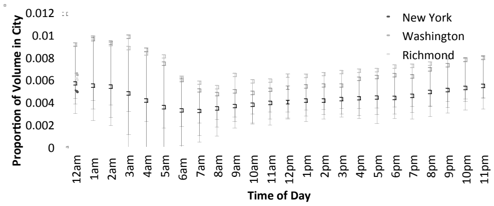
Figure 5: Volume of tweets for a city versus the number of keywords below the given thresholds

as we examine cities with volume lower than 5 million we see that few keywords have MAPE values lower than 50%. With such low numbers, we may not be able to detect reliable diurnal patterns, even for keywords that are expected to demonstrate them.

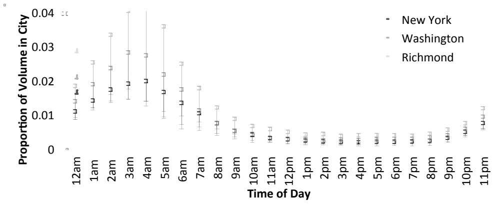
We now turn to address the questions about the connection between daily patterns and activities in cities. Can patterns capture real or virtual activities in cities? How much do the patterns represent daily urban routines? We first discuss results for three sample keywords. Then, we look at using keywords to find similarities across locations.

Figure shows the daily patterns of three sample keywords as a proportion of the total volume for each hour of the day. The keywords shown are "funny" in 6(a), "sleep" in 6(b) and "lunch" in 6(c). For each keyword, we show the daily curves in multiple cities: New York, Washington DC, and the much smaller Richmond, Virginia. The curve represents the $X24_{G,w}$ values for the keyword and location as described above. The error bars represent one standard deviation above and below the mean for each $X24$ series. For example, Figure 6(c) shows that in terms of proportion of tweets, the "lunch" keyword peaks in the three cities at 12pm. In Washington DC, for example, the peak demonstrates that 0.8% of the tweets posted between 12pm and 1pm include the word lunch, versus an average of about 0.05% of the tweets posted between midnight and 1am. The keywords were chosen to demonstrate weak ("funny") and strong ("lunch", "sleep") within-day variation across all locations. Indeed, the average entropy for "lunch" across the three locations is 3.95, the entropy for "sleep" is 4.1, while the "funny" entropy shows lower within-day variability at 4.55. The across-day variability is less consistent between the locations. For example, "sleep", the more popular keyword out of these three in all locations, has higher MAPE in New York than "funny", but lower MAPE in the other two locations. Figure 6(a) shows how the variability of "funny" rises (larger error bars) with the lower-volume cities.

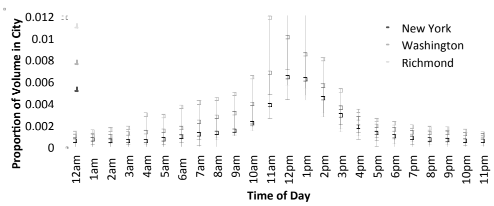
The figure demonstrates the mix between virtual and physical activities reflected in Twitter. The keywords "funny", and, for example, "lol" (the most popular keyword in our dataset) exhibit diurnal patterns that are quite robust for high-volume keywords and cities. However, those patterns are less stable and demonstrate high values of noise with lower volume for $X_{G,w}$. Moreover, the times in which they appear could be more reflective of the activities people perform online than reflect actual mood or happiness: 6am is perhaps not the ideal time to share funny videos. On the



(a) "Funny"



(b) "Sleep"



(c) "Lunch"

Figure 6: Daily patterns for three keywords, showing the average proportion of daily volume for each hour of the day

other hand, keywords that seem to represent real-world activities may provide a biased view. While it is feasible that Figure 6(c) reflects actual diurnal lunch patterns of the urban population, Figure 6(b) is not likely to be reflecting the time people go to sleep – peaking at 3am.

Nevertheless, we hypothesize that the diurnal patterns, representing virtual or physical activities, can help us develop an understanding of the similarities and differences in daily routines between cities. Can the similarity between "lunch" and "sleep" diurnal patterns in two cities suggest that the cities are similar?

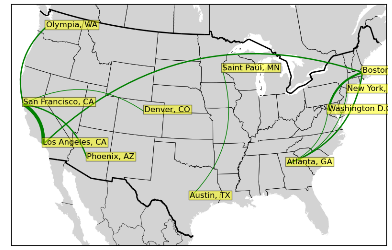
To initially explore this question, we compared the $X24$ series for a set of keywords, for each pair of cities. We measured the similarity of diurnal patterns using the Jenson-Shannon (JS) divergence between the $X24$ series of the same keyword in a pair of cities. JS is analogous to entropy, and defined over two distributions P and Q of equal size (in our case, 24 hours) as follows:

$$JS(P,Q) = \frac{1}{2}\left[\sum_{i=0}^{23} gp(i) + \sum_{i=0}^{23} gq(i)\right] \tag{3}$$

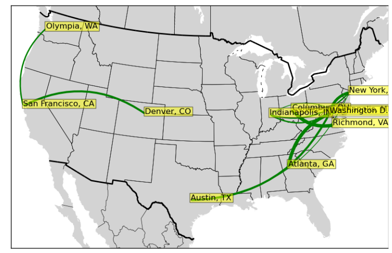
Where $gp(i) = \frac{p_i}{(p_i+q_i)/2}\log\frac{p_i}{(p_i+q_i)/2}$, and $gq(i) = \frac{q_i}{(p_i+q_i)/2}\log\frac{q_i}{(p_i+q_i)/2}$.

For our purposes, the distributions p_i and q_i were the normalized hourly mean-volume values for a specific keyword, as we did for the entropy calculation. The distance between two cities G_1, G_2 with respect to keyword w is thus $JS(p_{G_1,w}, p_{G_2,w})$. The distance with respect to a *set* of keywords is the sum over distances of the words in the set.

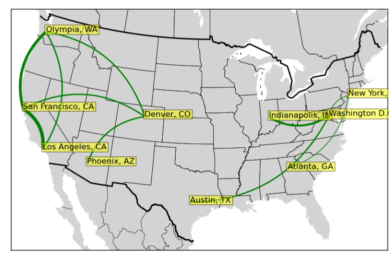
For comparison, we selected three sets of ten keywords to use when comparing locations: a random set of keywords, a



(a) Random



(b) Stable



(c) Significant

Figure 7: 10 most similar city pairs computed using different sets of keywords.

set of keyword with high entropy ("significant"), and a set of keywords with low MAPE ("stable"). We then plotted the top 10 most similar city pairs, according to each of these groups, on a map, as shown in Figure 7. The thickness of the arc represents the strength of the similarity (scaled linearly and rounded to integer thickness values from 1 to 5).

The similarities calculated using the 10 random keywords exhibit quite low coherence. While there are clearly a few city pairs that are closely connected (San Francsico-Los Angeles, Boston-Washington), for the most part it is hard to discern regular patterns in the data. Using the top 10 most stable keywords results in a more coherent cluster of major east-coast cities, along with a smaller, weakly connected cluster centering on San Francisco. When we calculate the similarity using the top 10 highest entropy keywords, the two clusters become much clearer, and the locality effect is stronger, resulting in a well-connected west-coast cluster centered around San Francisco, and an east-coast cluster in which Washington D.C. serves as a hub. Despite this promising result, it remains to be seen whether these effects are mostly due to timezone similarity between cities.

Discussion and Conclusions

We have extracted a large and relatively robust dataset of location data from Twitter, and used it to reason about diurnal pattern in different cities. Our method of collection of data, using the users' profile location field, is likely to have resulted in a high precision dataset with relatively high recall. We estimate that we are at least within an order of magnitude of the content generated in each city. While issues of Spam, errors and noise still exist, these issues were minimal (and largely ignored in this report).

This broad coverage allowed us to investigate the feasibility of a study of keyword-based diurnal patterns in different locations. We showed that patterns can be extracted for keywords, but that the amount of variability is highly dependent on the volume of activity for a city and a keyword. For low-volume cities, most keyword patterns are likely to be too noisy to reason about in any way. On the other hand, keywords with high within-day variability could be detected even for low-volume cities. The mapping from diurnal Twitter activities to physical-world (vs. virtual) activities and concepts is not yet well defined. Use of keywords like "funny" and "lol", for example, that can express mood and happiness (Golder and Macy 2011), can also relate to virtual activities people more readily engage with at different times of day. The reflection of physical activities can also be ambiguous, as shown with the analysis of the keyword "sleep" above.

This exploratory study points to a number of directions for future work. The variability error bars in Figure suggest that the $X_{G,w}$ series might be modelled well using a Poisson arrival model, which will apply itself better to studies and simulations. Topic models and keyword associations might help merge the data for a set of keywords to enhance the volume of data available for analysis for a given topic or theme. This method or others that could raise the volume of the analyzed time series can allow more robust analysis including, for example, detection of deviation from expected patterns for different topics.

Next steps will also include the development of better statistical tools for analyzing variability and stability of these time series, as well as comparisons against a model or other time series data. We especially are interested in comparing the diurnal patterns in social media to other sources of data – as social media data can augment or replace more costly methods of data collection and analysis.

Acknowledgments

This work was partially supported by the National Science Foundation CAREER-1054177 grant, a Yahoo! Faculty Engagement Award, and a Google Research Award. The authors thank Jake Hofman for his significant contributions to this work.

References

Ahern, S.; Naaman, M.; Nair, R.; and Yang, J. H.-I. 2007. World Explorer: visualizing aggregate data from unstructured text in geo-referenced collections. In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, 1–10. New York, NY, USA: ACM.

Cheng, Z.; Caverlee, J.; Lee, K.; and Sui, D. Z. 2011. Exploring Millions of Footprints in Location Sharing Services. In *ICWSM '11: Proceedings of The International AAAI Conference on Weblogs and Social Media*. AAAI.

Cheng, Z.; Caverlee, J.; and Lee, K. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*. New York, New York, USA: ACM Press.

Crandall, D.; Backstrom, L.; Huttenlocher, D.; and Kleinberg, J. 2009. Mapping the World's Photos. In *WWW '09: Proceeding of the 18th international conference on World Wide Web*. New York, NY, USA: ACM.

Cuff, D.; Hansen, M.; and Kang, J. 2008. Urban sensing: out of the woods. *Commun. ACM* 51(3):24–33.

Dodds, P. S.; Harris, K. D.; Kloumann, I. M.; Bliss, C. a.; and Danforth, C. M. 2011. Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. *PLoS ONE* 6(12):e26752.

Eisenstein, J.; O'Connor, B.; Smith, N. A.; and Xing, E. P. 2010. A latent variable model for geographic lexical variation. In *EMNLP '10: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1277–1287. Stroudsburg, PA, USA: Association for Computational Linguistics.

Girardin, F.; Fiore, F. D.; Ratti, C.; and Blat, J. 2008. Leveraging explicitly disclosed location information to understand tourist dynamics: a case study. *Journal of Location Based Services* 2(1):41–56.

Golder, S. A., and Macy, M. W. 2011. Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science* 333(6051):1878–1881.

Golder, S. A.; Wilkinson, D.; and Huberman, B. A. 2007. Rhythms of social interaction: messaging within a massive online network. In *CT '07: Proceedings of the 3rd International Conference on Communities and Technologies*.

Gonzalez, M. C.; Hidalgo, C. A.; and Barabasi, A.-L. 2008. Understanding individual human mobility patterns. *Nature* 453(7196):779–782.

Hecht, B.; Hong, L.; Suh, B.; and Chi, E. H. 2011. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *CHI '11: Proceedings of the 2011 annual conference on Human Factors in Computing Systems*, 237–246. New York, NY, USA: ACM.

Java, A.; Song, X.; Finin, T.; and Tseng, B. 2007. Why we Twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07: Proceedings of the WebKDD and SNA-KDD 2007 workshop on Web mining and social network analysis*, 56–65. New York, NY, USA: ACM.

Naaman, M.; Boase, J.; and Lai, C.-H. 2010. Is it really about me? Message content in social awareness streams. In *CSCW '10: Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, 189–192. New York, NY, USA: ACM.

Shimshoni, Y.; Efron, N.; and Matias, Y. 2009. On the Predictability of Search Trends. Technical report, Google. Available from http://research-.google.com/archive/google_trends_predictability.pdf.