Constructing a taxonomy of fine-grained human movement and activity motifs through social media

Morgan R. Frank[1,2], Jake Ryland Williams[1], Lewis Mitchell[1,3],
James P. Bagrow[1], Peter Sheridan Dodds[1], Christopher M. Danforth[1]

[1]Computational Story Lab, Department of Mathematics and Statistics,

Vermont Complex Systems Center, Vermont Advanced Computing Core,

University of Vermont, Burlington, Vermont, United States of America

[2]Current address:

Massachussetts Institute of Technology,

Cambridge, MA, USA

[3]Current address:

School of Mathematical Sciences,

University of Adelaide, Adelaide, Australia

mrfrank@mit.edu, jake.williams@uvm.edu, lewis.mitchell@uvm.edu,

james.bagrow@uvm.edu, peter.dodds@uvm.edu, chris.danforth@uvm.edu

May 15, 2015

Abstract

Profiting from the emergence of web-scale social data sets, numerous recent studies have systematically explored human mobility patterns over large populations and large time scales. Relatively little attention, however, has been paid to mobility and activity over smaller time-scales, such as a day. Here, we use Twitter to identify people's frequently visited locations along with their likely activities as a function of time of day and day of week, capitalizing on both the content and geolocation of messages. We subsequently characterize people's transition pattern motifs and demonstrate that spatial information is encoded in word choice.

1 Introduction

The growth of social media has made possible new insights into human behavior and predictability [1–5].

Many active areas of research, such as disease spreading [6–12], traffic forecasting [13,14], urban planning [15, 16], election prediction [31], understanding stock market behavior [32], and the spreading of ideas [33] have greatly benefited from the recent availability of large-scale social datasets, such as mobile phone data [17–21] and Twitter data [22–25]. One particular area of interest lies in discerning patterns of human activity, and understanding what these patterns tell us about human life. An example is the ability to identify the typical work day for members of western culture, who comprise most of the English speakers on Twitter [30].

Previous studies of macroscale human mobility [17, 22] have been bolstered by investigations into the mechanics of microscale activity. González et al. [20] used cellphone data to investigate daily mobility patterns, called "motifs", based on cellphone tower reception areas. Their study used data with very fine temporal resolution and coarse spatial resolution. In this study, we use geolocated Twitter data containing coarse temporal resolution due to subsampling, refined spatial resolution, and, crucially, the
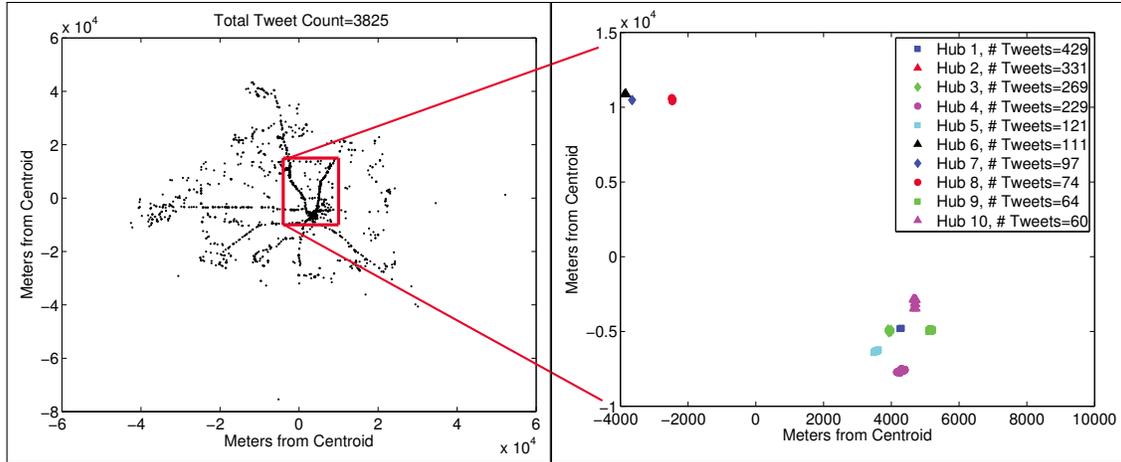
1

Figure 1: **Tweet hubs for an example prolific Twitter user. (Left) The raw locations for all of the geolocated tweets for this individual. The axes have been scaled to distance in meters from expected location to obfuscate the identity of the individual. (Right) A zoomed-in view of a subset of this individual's hubs of activity. Tweet locations are colored by hub, the legend list hubs ranked by frequency. Most hubs are positioned to the southeast of the centroid, while hubs 6, 7, and 8 appear in cluster to the northwest.**

lexical content of the Twitter messages [22]. These differences provide the opportunity for us to use new methods and sensing techniques for understanding human behavior through social data.

Twitter is a social media platform where individuals author short messages, called "tweets", which they can choose to label with their exact location. The combination of message content and spatial location is especially novel in comparison to mobile phone data. We have seen that changes in the culture of a society may be reflected in large text corpora, such as the Google n-gram dataset [34, 35]. Similarly, Twitter can provide large aggregated texts representing the underlying sentiment or well-being of a population of Twitter users [36, 37]. In previous work, we showed that Twitter word usage allows us to demonstrate the characteristics and similarities of American cities [23], and to explore the relationship between travel and happiness [22].

2   Methods

We collected a random ten percent of all Twitter activity through Twitter's garden hose API during 2011 and 2012. Many individuals choose to make their account public, and geolocate their messages from a GPS-enabled device, such as a smartphone. Approximately one percent of all Twitter activity is geolocated, and we collected roughly 150 million geolocated tweets during the time period of this study. Grouping the tweets by their sender, we iden-

tify "prolific" Twitter users as individuals with at least 600 geolocated tweets in our dataset. Given our sampling rate, we expect that these individuals each geolocated roughly 6000 messages during the 2 year period. We aim to investigate the daily pattern of life of these people. Prolific Twitter users in this study can be found all over the world, however the context analysis we present will reflect English speakers only. The entire user dataset is used to produce context-independent analysis. Prolific Twitter users are filtered to eliminate automated Twitter services, such as weather stations, emergency information, coupon services, job searching services, and even the Big Ben clock in London, England, by investigating repetitive message structure and repetition of key words. We obtained about $1,900$ prolific Twitter users for this study, of which $1,000$ are English speakers.

For each individual, we identify frequently visited spatial locations by finding geographic spatial clusters of geolocated tweets. We look for clusters containing at least 50 tweet locations such that the distance between any tweet in the cluster and the next closest tweet location in the cluster is at most 25 meters. We call such clusters "tweet hubs". Figure 1 demonstrates an example individual's raw geolocated tweets along with their tweet hubs. We find that slight perturbations in these parameters do not change our results appreciably.

We use results from the American Time Use Survey (ATUS) [38] conducted by the Bureau of Labor Statistics identifying activities by hour of the day in an effort to understand an individual's activity at a tweet hub. Par-
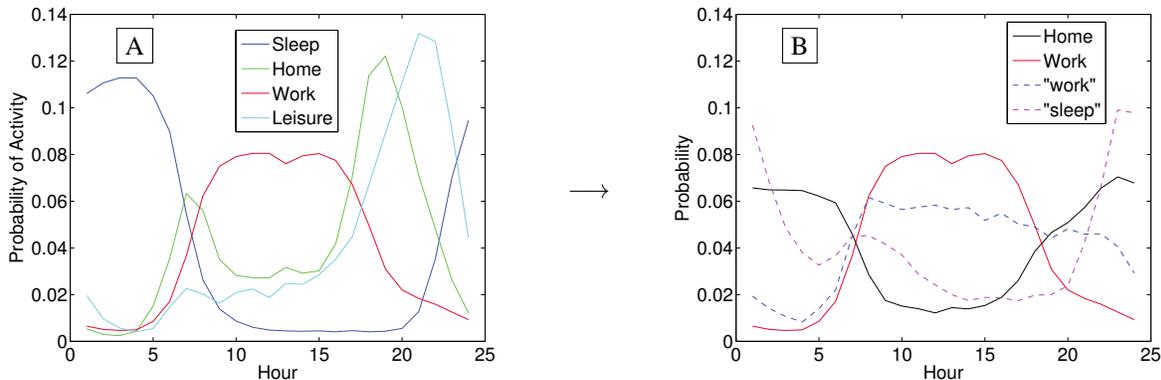
Figure 2: **(A) Probability density functions for each of the four activity classes in the American Time Use Survey [38]. (B) We combine the Sleep, Home, and Leisure curves from (A) to obtain a new "Home" hourly PDF representing non-working activity with a solid black line. The solid red line represents the Work curve. The dashed lines represent the hourly PDFs for occurrences of the words "work" and "sleep" on twitter during the years 2011 and 2012.**

ticipants in this survey were individuals age 15 and over who were employed full-time on days they worked. These data were collected by averaging survey results conducted in the years 2008 through 2012. The four activity classifications are Sleep, Home, Work, and Leisure, and Fig. 2A exhibits the hourly probability density functions (hourly PDFs) for each of those classes. The Sleep and Work classes are each clearly distinct from the hourly PDFs of any other activity class. Also, the hourly PDFs for the Home and Leisure classes are similar in the evening hours, but the curve representing the Home class exhibits a small peak in probability in the morning hours that is absent for the Leisure hourly PDF. Our preliminary investigations demonstrated that there were insufficient numbers of tweet hubs of the Sleep activity classification to merit meaningful results, and a contextual analysis of the vocabularies for tweet messages contained in Home or Leisure tweet hubs showed surprising similarities making them indistinguishable (details on these analyses below). Therefore, we combine the Sleep, Home, and Leisure activity curves shown in Figure 2A to produce the new Home hourly PDF shown in Figure 2B, which represents the union of the three separate activity classifications.

We classify each tweet hub as representing either a Home or Work activity by measuring the root mean square error (RMSE) between the tweet hub and ATUS PDFs. This method provides circumstantial evidence to explain why an individual frequents a particular area represented by a tweet hub. Figures 3A & 3B provide an example of this activity classification for tweet hubs from the individual in Fig. 1.

González et al. [20] viewed daily mobility patterns for individuals as networks based on their cellphone tower reception: nodes represent tower reception areas and directed edges represent travel between two reception areas. Inspired by this idea, we construct "cumulative Twitter motifs" as networks where each tweet hub of a given individual is represented as a node and weighted directed edges exists between node *A* and node *B* if the individual had a geolocated tweet contained in tweet hub *A* and then authored a tweet from hub *B* within a two hour time period. We call this a "transition" from tweet hub A to tweet hub B. We ignore transitions that do not occur at least five times in our dataset. This construction does not achieve as precise a measurement of daily mobility for the individuals as in [20], but nonetheless provides good approximations for the individuals' mobility patterns. This difference of methods is necessary because of the discrete temporal nature of the Twitter data. Our method of constructing cumulative Twitter motifs may result in networks with separate components (i.e., transitioning between locations in each of the components takes longer than two hours, or there is no data to provide evidence for a transition); we treat each of these components separately when investigating human mobility patterns. We present an example motif in Fig. 3C.

The tweet message content provides a unique insight into the users' lives that is absent in other human mobility datasets. Previous studies [22–25, 34–37] have demonstrated the usefulness of examining the vocabulary of large collections of tweets and other text sources. We demonstrate the predictive power of words used at a tweet hub by comparing binary decision trees [41–43] and Bayesian document classifiers [44, 45] as they predict the
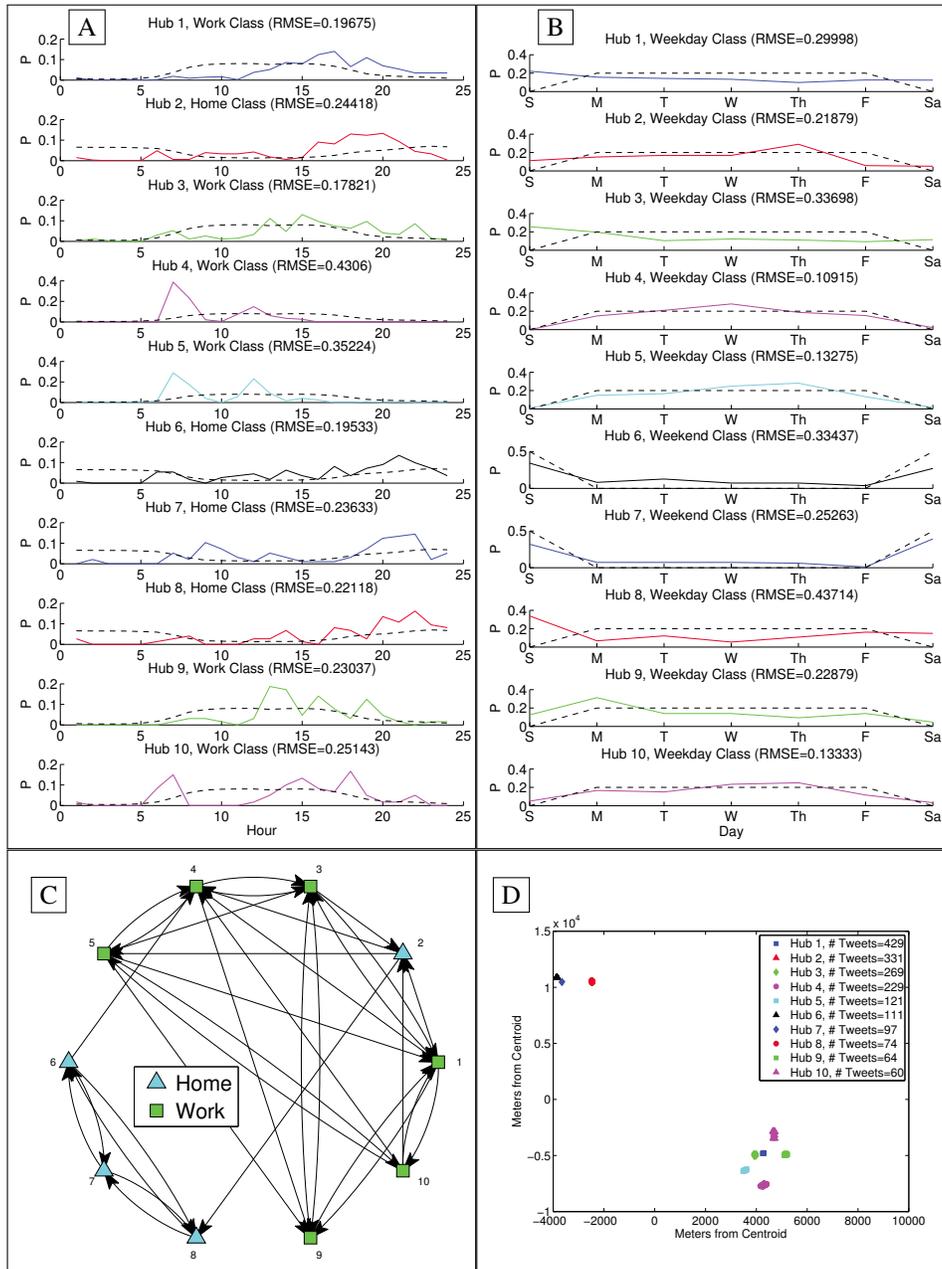
Figure 3: **An example Tweet hub classification and cumulative Twitter motif for the Twitter user from Fig. 1.** **(A) We classify the ten tweet hubs for this individual according to which of the activity class PDFs the hourly probability density function for that tweet hub is most similar to. For each of these tweet hubs, we provide the hourly PDF for that tweet hub in the color corresponding to the color of that tweet hub in Figure 1 in the right plot (note this same plot is provided in (D) for reference). The resulting hub classification is provided above each of the PDFs, and the PDF of the activity class is provided as a dashed line for comparison. We perform a similar treatment to obtain the daily classifications shown in panel (B). (C) The cumulative Twitter motif for the example individual with the nodes labeled by tweet hub and colored by the activity classification assigned to that tweet hub. Home-activity nodes 6, 7, & 8 are nearly their own network component, and this separation from the rest of the tweet hubs is reflected in the spatial distance shown in (D).**

4

activity classifications of tweet hubs in our dataset using word counts for the 50,000 most frequently used words in the English language according to Google Books, music lyrics, New York Times articles, and Twitter messages [24]. For each of these words, we obtain a word count representing the number of occurrences of that word in a text. As a baseline, we will note the improvement in the fraction of correct activity class predictions yielded from a null predictor, which is obtained by randomly permuting the true tweet hub classifications in our dataset and noting which classifications remain correct. The probability of a correct prediction under this process is given by

$$P(\text{correct}) = P(\text{Home})^2 + P(\text{Work})^2, \qquad (1)$$

where $P(\text{Home}) = 0.65$ and $P(\text{Work}) = 0.35$ according to tweet hubs taken from English speaking Twitter users. This results in a binomial distribution with a mean of $P(\text{correct}) = 0.545$ and a variance of $P(\text{correct})(1 - P(\text{correct})) = 0.248$. We could have naively assumed each activity classification is equally likely to occur, but the comparison to the more sophisticated null model is more realistic.

We randomly select half of the tweet hubs in our dataset as training data for a binary decision tree. The remaining tweet hubs are our validation dataset. To construct the binary decision tree, we first examine every possible binary split on each of the 10,000 word counts. We select the split that minimizes Gini's index [39,40], which measures the diversity of the resulting classification predictions. We represent the word count and the binary splitting criterion as a node on the decision tree with two child nodes. We recursively repeat this procedure for the children nodes until a node is found to yield predictions of only one class type according to the training data. Nodes with no children are called "leaf nodes". It is often found that binary decision trees over-fit the training data during construction, and improvements in prediction on validation data can be obtained by systematically removing, or "pruning", leaf nodes and nodes that are below a certain depth. We select the level of pruning that maximizes the fraction of correct predictions on the validation data.

In training the Bayesian document classifiers, we randomly select half of the tweet hubs to be training data, and we leave the remaining tweet hubs for validation. We restrict to $N = 1,000$ most frequently occurring words in tweet hubs from the training dataset. For each activity classification, we combine and normalize the word counts from tweet hubs of that class in the training data to obtain a likelihood of the word occurring given the activity class. We denote this by $P(w_i \mid c)$ where w_i is the ith word for $i \in \{1, \ldots, 1000\}$, and $c \in \{\text{Home}, \text{Work}\}$ is the activity

classification. Thus, given a tweet hub, H, the posterior probability that H is of class $c_H = c$ is given by

$$P(c_H = c) = P(c) \cdot \prod_{i=1}^{N} P(w_i \mid c)^{f_H(w_i)}, \qquad (2)$$

where $P(c)$ is the probability of activity classification c and $f_H(w_i)$ is the frequency of word w_i occurring in tweet hub H. The activity classification with greatest posterior probability is selected as the predicted activity classification for H.

Calculating the posterior probabilities in Eq. 2 directly can lead to large computer roundoff errors. Thus, we instead consider the log probability when implementing the classifier. This allows us to frame the calculation in terms of minimizing surprisal as opposed to maximizing likelihood. Furthermore, since the two activity classifications in our dataset exist in roughly equal proportions, the majority of variation in the posterior probabilities is due to the likelihood function and the performance of the Bayesian document classifier is not appreciably diminished by assuming a flat prior probability distribution, i.e., $P(\text{Home}) = P(\text{Work})$. Accordingly, Eq. 2 becomes

$$-\log P(c_H = c) = -\log(2) - \sum_{i=1}^{N} f_H(w_i) \cdot \log P(w_i \mid c). \tag{3}$$

Equation (3) provides an interesting opportunity to determine which words discriminate most strongly between the two activity classifications. We examine the difference in the logs of the two posterior probabilities, P_{comp}, for a tweet hub H, to find

$$P_{\text{comp}} = \sum_{i=1}^{N} f_H(w_i) \Big(\log P(w_i \mid \text{Home}) - \log P(w_i \mid \text{Work}) \Big). \tag{4}$$

We can rank words by their contribution to the total by examining the sum term by term. This ranking allows us to see the contextual differences in the vocabularies of the Home and Work activity classifications.

3 Results

We first investigate global statistics after identifying tweet hubs for each of the prolific Twitter users. Figure 4A shows the probability an individual has N tweet hubs. As was found in [20], we observe a trend that is approximately lognormal and is well modeled by

$$F(N; \mu, \sigma) = \frac{\exp\left(-\frac{(\ln N - \mu)^2}{2\sigma^2}\right)}{\sqrt{2\pi} N \sigma}, \tag{5}$$
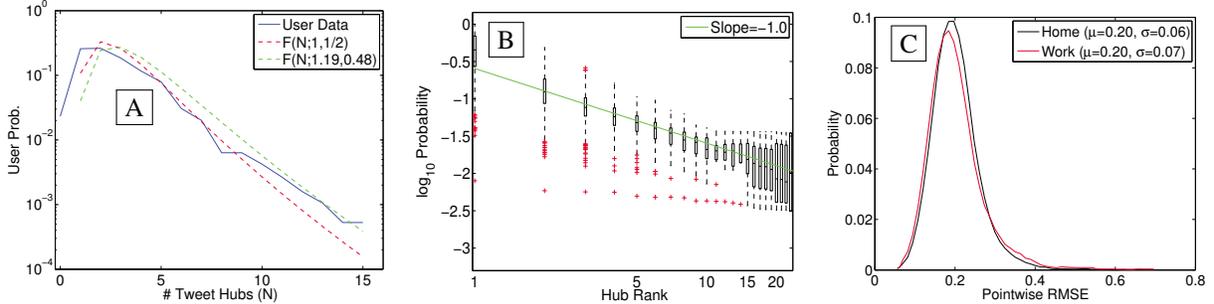
5

Figure 4: **(A) A semi-logarithmic plot demonstrating the lognormal relationship between the number of tweet hubs (N) to the probability of individuals with that many tweet hubs. As was found in [20], we see that the relationship is well modeled by $F(N) = \exp\left(-\frac{(\ln N - \mu)^2}{2\sigma^2}\right)/(\sqrt{2\pi}N\sigma)$ with $\mu = 1$ and $\sigma = 1/2$ (red dashed line); however, we find that the model fits our data slightly better in terms of least squares with $\mu \approx 1.19$ and $\sigma \approx 0.48$ (green dashed line). (B) After ranking the tweet hubs of each individual by tweet count, we compare tweet hub rank to the probability of the individual of that tweet hub tweeting from that tweet hub. For each hub rank, we provide a box-plot of the probabilities of tweets occurring in tweet hubs of that rank. Outliers are represented by red crosses. We provide a green solid line with slope -1 as a guide for the eye. The box-plots for lower rank-ing tweet hubs represent many more data points than the box-plots for the higher ranking tweet hubs because only a few individuals were found to have more than 10 tweet hubs (see A). (C) The distributions for RMSE for the tweet hub hourly PDFs compared to the activity PDF they were most similar to.**

where $\mu = 1$, and $\sigma = 1/2$; however, a slightly better fit in the least squares sense is obtained when $\mu = 1.19$ and $\sigma = 0.48$. This finding supports the "universal law" proposed by González et al. in [20]. Figure 4B demonstrates the relationship between ranking tweet hubs by tweet count and the probability that the individual sent a tweet from that hub. We provide a line with slope -1 as a guide for the eye. This linear trend may represent some underlying mechanism, but we do not endeavor to explore this further here.

We found that 68% (65% for English speakers) of tweet hubs in our dataset were classified as Home, while 32% (35% for English speakers) were classified as Work. Figure 4C shows us how similar hourly PDFs for tweet hubs in our dataset were when compared to the activity class PDF (Fig. 2) the tweet hub was classified as. We find that the root mean square error when comparing the hourly PDFs for tweet hubs and their activity classifications is almost always less than 0.4, indicating that our classification method is performing reasonably well in terms of PDF similarity.

We investigate the cumulative Twitter motifs in Fig. 5, where we provide networks for each motif component that we encountered at least 1% of the time. Each directed edge in these networks is labeled to show what fraction of transitions between tweet hubs in that component is explained by that edge. The nodes also have subplots displaying the probabilities for the tweet hub to have each of the activity classifications. The final plot shows the probabilities of observing each of the prominent motif components.

We first notice that over 60% of cumulative network components were found to be an isolated point or a 2-cycle (i.e. the Rank 1 or Rank 2 networks). Interestingly, one of the nodes in the 2-cycle is more likely to represent a Work tweet hub than the other and the edge originating from this node represents almost two-thirds of the transitions between these tweet hubs. This reflects the finding in [22, 24] that the volume of Twitter activity is increased in the evening when a working Twitter user would be commuting from work or school to home.

We next note that the rank 4 through 8 motif components represent three hubs. The Rank 4, 5, 7, & 8 networks suggest that nodes that are more likely to be Home than Work tend to have higher degrees, and the sum of weights from incoming edges tends to outweigh the sum of weights from outgoing edges. These observations suggest these nodes play a central role in the mobility patterns represented by these networks.

We capitalize on the content of the Twitter messages by comparing binary decision trees and Bayesian document classifiers in predicting the activity classifications of tweet hubs based only on the vocabulary of the tweets contained in those hubs. In Fig. 6, we compare the fraction of correct classification predictions with the null model. Points to the right of the black vertical dashed
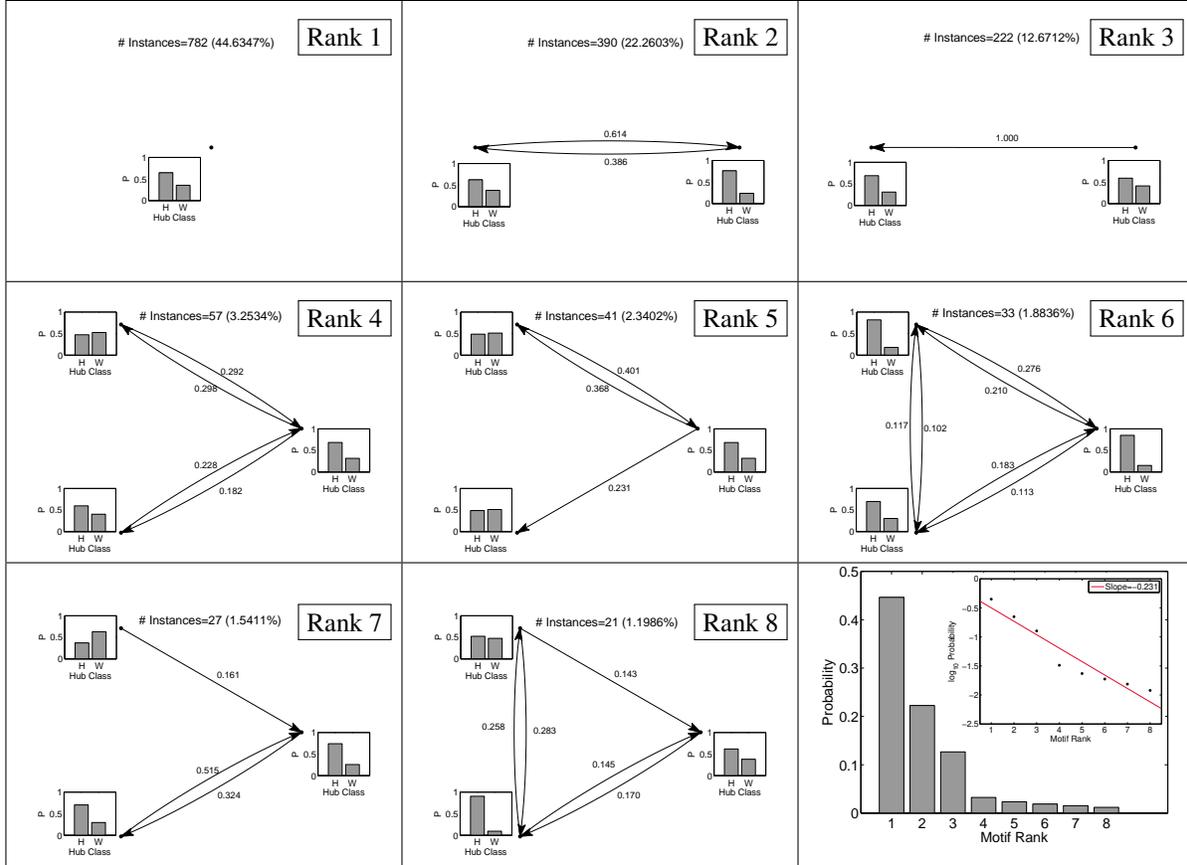
Figure 5: **We construct aggregated Twitter mobility motifs for each prolific user as directed networks, where tweet hubs for that user are represented by nodes, and a directed edge from node A to node B exists if a tweet was sent from tweet hub A and another tweet was sent from tweet hub B at most two hours apart. We separate each of these networks into their components and group these isomorphic subnetworks while counting how many of each transition between nodes occurs and the activity classification types of each of the nodes. We rank these subnetworks by occurrences among the prolific users and display the most common subnetworks in panels 1-8. The edges are labelled to convey the fraction of observed transitions represented by that edge. The nodes each have a subplot showing the probabilities that the tweet hub represented by that node has each of the activity classifications (H=Home, W=Work). The last panel shows the probabilities of these subnetworks in rank-order, along with a subplot containing the same distribution on a log-scale.**

line represent predictions that performed better than the null model distribution with a p-value$< 10^{-3}$. For the binary decision tree and the Bayesian document classifier, we randomly select half of the prolific Twitter users in our dataset as training data and we record the percent of the correct activity classifications on the remaining half of the users as validation. Repeating this procedure for 3,000 trials, we obtain the blue and green distributions, representing the predictive power of the binary decision tree and the Bayesian document classifier, respectively.

Both prediction methods yield a proportion of cor-

rect classification predictions that perform better than expected from the null model (i.e. p-value$< 10^{-3}$). On average, the null model yielded correct predictions 54.5% of the time, the binary decision tree yielded correct predictions 59.6% of the time, and the Bayesian document classifier yielded correct predictions 60.7% of the time. Throughout several trials of training and validating our predictors, we found that 15% of Home tweet hubs and 20% of Work tweet hubs were correctly classified at least 99% of the time, indicating classification predictions were not random and not biased towards one class despite the
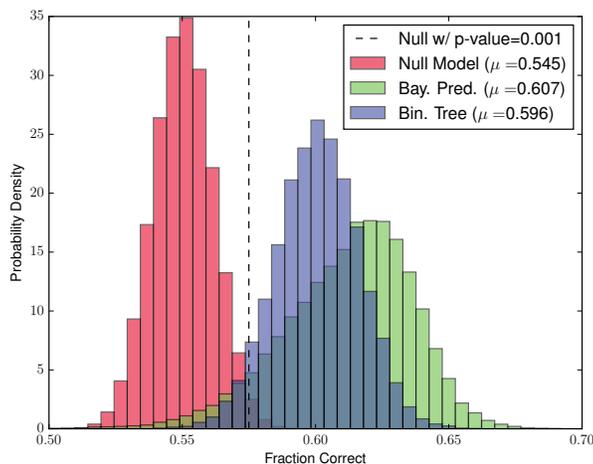
Figure 6: Distributions of the fraction of correct predictions made using three approaches: the null model (red); a pruned binary decision tree (blue); and a Bayesian document classifier (green). The latter two distributions are both estimated from 3,000 trials (see main text), and a "fraction correct" calculated according to the number of correct predictions divided by the total number of tweet hubs in the validation dataset of that trial. The dashed line represents the value of the null model distribution above which has a p-value $< 10^{-3}$.

different number of hubs of each class (see Fig. 4C). Furthermore, both of these predictive methods allow us to see which words contribute the most to the predictions. We provide the words that most strongly discriminate between Home (Fig. 7A) and Work (Fig. 7B). We see that pronouns (e.g. "i", "you", "she"), slang (e.g. "lol", "lmfao", "gonna"), and profanity (e.g. "a**", "b**ch") are more likely to occur from a Home tweet hub than from a Work tweet hub. On the other hand, references to time of day (e.g. "lunch", "afternoon", "goodmorning"), references to weekdays, and references to work (e.g. "business", "working") are more likely to occur at Work tweet hubs.

4 Discussion

The purpose of this study is to use geolocated Twitter data to understand and characterize aspects of daily human life, such as identifying key spatial locations, how people travel between their locations, and why they visit those locations. Our work improves our understanding of sociotechnical systems in a way that may improve traffic models, models of disease spread, and current methods of urban planning by introducing methods for utilizing large-scale Twitter data that is abundant and publicly available. Using Twitter data for understanding sociotechnical systems grants us the interesting opportunity to use the content in the tweet messages to both asses our methods of

characterizing human activity at key locations and infer the activities of individuals at locations with more incomplete data.

We find evidence in our Twitter data to support observations on human mobility made through other data sources. For example, Fig 4A exhibits a lognormal distribution of tweet hub counts that agrees the finding of González et al. [20], which employed both survey and cellphone data. Furthermore, we find agreement with previous studies finding that the mobility patterns of most people are explained by just three locations, and tweet hubs representing Home activity are more likely to play a centralized role in the mobility pattern (see Fig. 4A & Fig. 5). Finally, we find that we are more able to capture data from commutes from Work locations to Home locations, which may be related to previous findings regarding increased tweet volume in the latter portion of each day [22, 36]. These conclusions may inform future studies by reducing the degrees of freedom required to model human mobility dynamics.

We have introduced novel methods for observing characteristics of the daily lives of prolific Twitter users. Providing a way to identify key locations for individual users as tweet hubs allows us to analyze their mobility between those spatial locations, as well as attempt to understand why they visit those locations. We demonstrate our methods' ability to achieve the binary classification of tweet hubs between Home and Work by observing a

Figure 7: **Words ranked by their percent contributions to P_{comp} as terms in the sum from Eq. (4) for tweet hubs that were consistently classified as Home (A), or Work (B).**

notable subset of tweet hubs with activity classifications that are correctly identified almost all of the time, despite the difference in sample sizes among the Home and Work tweet hubs and regardless of what subset of our data we use to train our methods (see Fig. 6). Furthermore, an analysis of the distinguishing words for the two activity classes for the consistently correctly classified hubs supports our classifications by demonstrating the use of slang, profanity, and pronouns at Home, while tweets sent from Work hubs contained more formal grammar with references to work, time of the day, and day of the week (see Fig. 7).

Finally, our use of the words contained in tweets from hubs introduces the exciting power of the content at a location to aid in identifying an individual's activity at that location. Fig. 7 provides evidence that just the vocabulary of tweets sent from a location allows us to accurately predict the activities of the individual at that location. A more detailed understanding of language along with more sophisticated computational methods will allow researchers to identify the activities of an individual at a key location with greater sensitivity, beyond a binary classification.

References

[1] R.W. Lariscy, E.J. Avery, K.D. Sweetser, P. Howes. *An examination of the role of online social media in journalists source mix.* Public Relations Review, Volume 35, Issue 3, September 2009, Pages 314316 http://dx.doi.org/10.1016/j.pubrev.2009.05.008

[2] B.G. Smith. *Socially distributing public relations: Twitter, Haiti, and interactivity in social media Public Relations Review*, Volume 36, Issue 4, November 2010, Pages 329335. http://dx.doi.org/10.1016/j.pubrev.2010.08.005

[3] K.D. Sweetser, T. Kelleher. *A survey of social media use, motivation and leadership among public relations practitioners.* Public Relations Review, Volume 37, Issue 4, November 2011, Pages 425428. http://dx.doi.org/10.1016/j.pubrev.2011.08.010

[4] E. Fischer, A.R. Reuber. *Social interaction via new social media: (How) can interactions*

9

on Twitter affect effectual thinking and behavior?. Journal of Business Venturing, Volume 26, Issue 1, January 2011, Pages 118. http://dx.doi.org/10.1016/j.jbusvent.2010.09.002

[5] A.S. Huberman (2010) *Predicting the future with social media*. pp. 492499.

[6] V. Colizza, A. Barrat , M. Barthelemy, A.-J. Valleron & A. Vespignani. *Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions*. PLoS Medicine 4, 95110 (2007).

[7] S. Eubank et al. *Controlling epidemics in realistic urban social networks*. Nature 429, 180184 (2004).

[8] L. Hufnagel, D. Brockmann, & T. Geisel. *Forecast and control of epidemics in a globalized world*. Proc. Natl Acad. Sci. USA 101, 1512415129 (2004).

[9] C. Viboud, O.N. Bjomstad, D.L. Smith, L. Simonsen, M.A. Miller, B.T. Grenfell. 2006. *Synchrony, waves, and spatial hierarchies in the spread of influenza*. Science 312, 447-451. (doi:10.1126/science.1125237)

[10] N.M. Ferguson, D.A.T. Cummings, C. Fraser, J.C. Cajka, P.C. Cooley , D.S. Burke. 2006. *Strategies for mitigating an influenza pandemic*. Nature 442, 448-452. (doi:10.1038/nature04795)

[11] M.L. Ciofi degli Atti, S. Merler, C. Rizzo, M. Ajelli, M. Massari, P. Manfredi, C. Furlanello, G.S. Tomba, M. Iannelli. (2008) *Mitigation measures for pandemic influenza in Italy: an individual based model considering different scenarios*. PLoS ONE 3, e1790. (doi:10.1371/journal.pone.0001790)

[12] M. Ajelli, B. Gon calves, D. Balcan, V. Colizza, H. Hu, J.J. Ramasco, S. Merler, A. Vespignani. (2010) *Comparing large-scale computational approaches to epidemic modeling: agent-based versus structured metapopulation models*. BMC Infect. Dis. 10, 190. (doi:10.1186/1471-2334-10-190)

[13] R. Schlich & K.W. Axhausen. *Habitual travel behaviour: evidence from a six-week travel diary*. Transportation 30, 1336 (2003).

[14] R. Kitamura, C. Chen, R.M. Pendyala, & R. Narayaran. *Micro-simulation of daily activity-travelpatterns for traveldemand forecasting*.Transportation27,2551 (2000).

[15] M.W. Horner & M.E.S. OKelly. *Embedding economies of scale concepts for hub networks design*. J. Transp. Geogr. 9, 255265 (2001).

[16] M.C. González, P.G. Lind, & H.J. Herrmann. *A system of mobile agents to model social networks*. Phys. Rev. Lett. 96, 088702 (2006).

[17] M.C. González, C.A. Hidalgo, & A.L. Barabási. *Understanding individual human mobility patterns*. Nature, Vol. 453, No. 7196. (2008), pp. 779-782, doi:10.1038/nature06958

[18] J.P. Bagrow & Y.-R. Lin. *Mesoscopic Structure and Social Aspects of Human Mobility*. PLoS ONE, 7 (5): e37676 (2012)

[19] C.A. Hidalgo & C. Rodriguez-Sickert. *The dynamics of a mobile phone network*. Physica A 387, 30173024 (2008).

[20] C.M. Schneider, V. Belik, T. Couronné, Z. Smoreda, & M.C. González. *Unravelling daily human mobility motifs*. Journal of The Royal Society 10 (84). doi:10.1098/?rsif.2013.0246

[21] D. Wang, D. Pedreschi, C. Song, F. Gionatti, & A.-L. Barabási. (2011) *Human mobility, social ties, and link prediction*. In: KDD, pp 11001108

[22] M.R. Frank, L. Mitchell, P.S. Dodds, C.M. Danforth. *Happiness and the Patterns of Life: A Study of Geolocated Tweets*. Nature Scientific Reports. (2013) Macmillan Publishers LiMITed. http://dx.doi.org/10.1038/srep02625

[23] L. Mitchell, K.D. Harris, M.R. Frank, P.S. Dodds, C.M. Danforth. *The Geography of Happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place*. PLoS ONE. http://www.plosone.org/article/info:doi/10.1371/journal.pone.0064417

[24] P.S. Dodds, K.D. Harris, I.M. Kloumann, C.A. Bliss, C.M. Danforth. (2011) *Temporal Patterns of Happiness and Information in a Global-Scale Social Network: Hedonometrics and Twitter*. PLoS ONE 6(12): e26752. doi:10.1371/journal.pone.0026752

[25] C.A. Bliss, I.M. Kloumann, K.D. Harris, C.M. Danforth, P.S. Dodds. (2012) *Twitter reciprocal reply networks exhibit assortativity with respect to happiness*. Journal of Computational Science 3(5), pp. 388-397, http://dx.doi.org/10.1016/j.jocs.2012.05.001

[26] A.-L. Barabási. *The origin of bursts and heavy tails in human dynamics*. Nature 435, 207211 (2005).

[27] D. Brockmann, L. Hufnagel, & T. Geisel. *The scaling laws of human travel*. Nature 439, 462465 (2006).

[28] G. Ramos-Fernandez et al. *Lévy walk patterns in the foraging movements of spider monkeys* (Ateles geoffroyi). Behav. Ecol. Sociobiol. 273, 17431750 (2004)

[29] J. Klafter, M.F. Shlesinger, & G. Zumofen. *Beyond brownian motion*. Phys. Today 49, 3339 (1996).

[30] http://www.peerreach.com

[31] Tumasjan, Andranik, et al. *Predicting elections with twitter: What 140 characters reveal about political sentiment.* Proceedings of the fourth international international conference on weblogs and social media. 2010.

[32] Kirilenko, Andrei, et al. *The flash crash: The impact of high frequency trading on an electronic market.* Available at SSRN 1686004 (2011).

[33] Peter Hedstrom. *Experimental Macro Sociology: Predicting the Next Best Seller*. Science 10 February 2006: 311 (5762), 786-787. [DOI:10.1126/science.1124707]

[34] J.B. Michel, Y.K. Shen, A.P. Aiden, A. Veres, M.K. Gray, et al. (2011) *Quantitative analysis of culture using millions of digitized books*. Science Magazine 331: 176182.

[35] Google Labs ngram viewer. Availabe at http://ngrams.googlelabs.com/. Accessed October 24, 2011.

[36] P. S. Dodds & C.M. Danforth. *Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents*. Journal of Happiness Studies, (2009). dpi: 10.1007/s10902-009-9150-9

[37] I.M. Kloumann, C.M. Danforth, K.D. Harris, C.A. Bliss, & P.S. Dodds. (2012) *Positivity of the English Language.* PLoS ONE 7(1): e29484. doi:10.1371/journal.pone.0029484

[38] Bureau of Labor Statistics, American Time Use Survey.

[39] C. Gini. (1912). "Variabilit e mutabilit" 'Variability and Mutability', C. Cuppini, Bologna, 156 pages. Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi (1955).

[40] C. Gini. (1909). "Concentration and dependency ratios" (in Italian). English translation in Rivista di Politica Economica, 87 (1997), 769789.

[41] Hyafil, Laurent, and R.L. Rivest. *Constructing optimal binary decision trees is NP-complete.* Information Processing Letters 5.1 (1976): 15-17.

[42] E.M. Rounds. *A combined nonparametric approach to feature selection and binary decision tree design.* Pattern Recognition 12.5 (1980): 313-317.

[43] S. Shlien. *Multiple binary decision tree classifiers.* Pattern Recognition 23.7 (1990): 757-763.

[44] A. McCallum, & K. Nigam. *A comparison of event models for naive bayes text classification.* AAAI-98 workshop on learning for text categorization. Vol. 752. 1998.

[45] D.D. Lewis. *Naive (Bayes) at forty: The independence assumption in information retrieval.* Machine learning: ECML-98. Springer Berlin Heidelberg, 1998. 4-15.