Environment and Planning B: Urban Analytics and City Science
XX(X):1–19
©The Author(s) 2017
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

A year in Madrid as described through the analysis of geotagged Twitter data

Travis R. Meyer[1], Daniel Balagué[14], Miguel Camacho-Collados[15], Hao Li[1], Katie Khuu[2], P. Jeffrey Brantingham[3], and Andrea L. Bertozzi[1]

Abstract

Gaining a complete picture of the activity in a city using vast data sources is challenging yet potentially very valuable. One such source of data is Twitter which generates millions of short spatio-temporally localized messages that, as a collection, have information on city regions and many forms of city activity. The quantity of data, however, necessitates summarization in a way that makes consumption by an observer efficient, accurate, and comprehensive. We present a two-step process for analyzing geotagged twitter data within a localized urban environment. The first step involves an efficient form of latent Dirichlet allocation (LDA), using an expectation maximization, for topic content summarization of the text information in the tweets. The second step involves spatial and temporal analysis of information within each topic using two complimentary metrics. These proposed metrics characterize the distributional properties of tweets in time and space for all topics. We integrate the second step into a graphical user interface that enables the user to adeptly navigate through the space of hundreds of topics. We present results of a case study of the city of Madrid, Spain, for the year 2011 in which both large-scale protests and elections occurred. Our data analysis methods identify these important events, as well as other classes of more mundane routine activity and their associated locations in Madrid.

Keywords

Urban sensing, topic model, Latent Dirichlet Allocation, spatial analysis

[1] Dept. of Mathematics, Univ. of California Los Angeles, USA
[2] Dept. of Computer Science, Univ. of California, Irvine, USA
[3] Department of Anthropology, Univ. of California Los Angeles, USA
[4] Research Computing and Cyberinfrstructure; Department of Mathematics, Applied Mathematics and Statistics, Case Western Reserve University, USA
[5] Spanish National Police Corps Madrid; Dept. of Statistics and Operations Research, Univ. of Granada, Spain

Corresponding author:
P. Jeffrey Brantingham, Department of Anthropology, Univ. of California Los Angeles, USA
Email: branting@.ucla.edu

Introduction

Human dynamics within cities are complex and constantly changing. Emergent activities and behavioral trends can be difficult to characterize in a precise way because cities are composed of a multitude of people interacting in complex ways against a heterogeneous social and physical backdrop. Starting in 2006, the service known as Twitter [*] began a rapid rise to global popularity and use. Twitter enables anyone with an appropriate device to generate a brief message of 140 characters or fewer and post it to the service for other users to see. Twitter later enabled users to specify not only the message, but a location described by latitude and longitude with each post. This location, known as a geotag, allows further analysis to be performed relating the content of the posts, known as tweets, to specific areas of the city where they were generated. In this paper we present a series of computationally efficient steps to summarize millions of geotagged tweets from an urban environment to provide a concise portrait over space and time that meaningfully represents activity within the city. We use the year 2011 in Madrid as a case study. Such a tool has the potential to impact many organizations from local government including city planning, policing and crime analysis, to scientific studies such as social analysis of protests and gatherings, to more broad based urban policy and planning applications.

Twitter has featured in many previous investigations due to the richness of the information content. A significant amount of analysis can be performed on tweets without location information, for example with keyword and keyphrase extraction using probabilistic models (Zhao et al., 2011), or modeling the full structure of tweets in Twitter conversations (Ritter et al., 2010). The interaction between users can be analysed to predict “re-tweeting” behavior (when a user broadcasts another user’s tweet) (Suh et al., 2010), or to find important users responsible for influencing the twitter network (Laflin et al., 2012). With location information, many other lines of investigation are available. Modeling tweet content and locations simultaneously can find latent factors that relate specific users to the use of words at particular locations (Hu et al., 2013). This task of predicting user locations is a popular area of work, with recent efforts focusing on tweet content analysis (Han et al., 2014) and propagation of location information from a few geolocated users to all users via the network social structure (Compton et al., 2014). In contrast to these methods, we are interested in using a content model and tweet location to study structure of the city, rather than social structure linking Twitter users.

Prior work most closely related to our contribution uses social media for urban mapping and detecting behavioral trends (Kling and Pozdnoukhov, 2012; Frias-Martinez and Frias-Martinez, 2014), crime prediction (Gerber, 2014), police interaction with the community (Heverin and Zach, 2010), event prediction (Asur and Huberman, 2010) and detection (Sakaki et al., 2010; Weng and Lee, 2011), and human behavior analysis (Farrahi and Gatica-Perez, 2011). Twitter data are often processed using variations of latent Dirichlet allocation (LDA) which presents their own unique difficulties as discussed in Hong and Davison (2010). Nevertheless LDA provides a sound basis for understanding Twitter content (Ramage et al., 2010). In Hong et al. (2012) LDA is modified to specifically find geographical locations associated with topics; however, this is not applied at the level of a city for studying behavior therein. This work demonstrates that social media content, including that of Twitter, is capable of providing detailed information about the structure and dynamics of activity in cities. For example, Kling and Pozdnoukhov (2012) analyze daily temporal dynamics of topics in New York. The current work is similar in the sense

[*] http://www.twitter.com/

that we utilize topic modeling to study tweets and to build an understanding of city activity patterns in both space and time. Unlike Kling and Pozdnoukhov (2012) and Farrahi and Gatica-Perez (2011), we investigate activity over a full year rather than daily or weekly patterns.

In the present work, we apply a recent algorithm for LDA (Asuncion et al., 2009) that requires an order of magnitude less computation time than methods based on the Gibbs sampler, when applied to large volumes of data. Our raw data for 2011 are comprised of approximately 1.4 million geotagged tweets and is reduced to 300 topics by this method. To sort through hundreds of topics, we propose four metrics, two in space and two in time, that quantify properties of the topic distributions throughout the city. These metrics quantify the degree to which spatial and temporal patterns are unimodal-multimodal and concentrated-diffuse. To better navigate this information, we have developed a tool in the form of a graphical user interface (GUI)[†] that enables a user to quickly browse and consume analysis results, a field of work that is in need of development (Schwartz et al., 2013).

We then combine these tools to perform an original in-depth look into tweets from the city of Madrid in the year 2011. This year saw significant political activity in the form of anti-austerity movements that ultimately triggered similar global movements, as well as elections later in the year. We demonstrate how these events, in addition to other city activity, are identified by the analytical methods discussed here. We show that many topics can be characterized by their metrics in time and space. When combined with a study of topic content, they provide insights into the activities taking place within Madrid.

In the following sections, we first present the topic modeling algorithm used to understand the content of the vast collection of text available using Twitter. This is followed by a presentation of the proposed metrics for characterizing the spatial and temporal distributions of topics. Finally, the result of applying these tools to the year 2011 in the city of Madrid, Spain is covered in-depth before final concluding remarks.

Latent Dirichlet Allocation

Tweets contain significant information about a city that is obfuscated in the casual language used. In addition, to study millions of tweets by hand is an unreasonable task. We seek to condense and clean this information using an approach known as latent Dirichlet allocation (LDA) (Blei et al., 2003). LDA is a probabilistic generative model for documents, or tweets, based on sampling words from distributions in a dictionary. LDA can be used to generate a “soft” unsupervised clustering of tweets by assigning to each word in each tweet a topic the word belongs to. Each tweet’s content is modeled with a distribution over t latent topics. Each topic, in turn, is a distribution over words in the vocabulary. LDA is related to mixture models (Ding et al., 2008) such as non-negative matrix factorization (Lee and Seung, 1999) and probabilistic latent semantic indexing (Hofmann, 1999). LDA differs from these models by regularizing the topic distributions (Blei et al., 2003) with a Dirichlet prior. This prevents zero entries that result in sub-optimal solutions.

LDA works as follows: we have m unique words used to compose n tweets. Let $\mathbf{A}_{i,j}$ be the number of times that the i^{th} word appears in the j^{th} tweet. This sparse matrix is the word-count matrix. Notably it discards word ordering information hence treating tweets as unordered bags of words. Assume there exist

[†] http://paleo.sscnet.ucla.edu/TwitterGUI.zip

m-by-t and t-by-n non-negative matrices $\mathbf{W}$ and $\mathbf{H}$ with unit sum columns such that $\mathbf{A}$ was generated as follows:

$$\begin{aligned} \mathbf{W}_{:,k} &\sim \text{Dir}(\alpha) \quad \forall\, k = 1, ..., t \\ \mathbf{H}_{:,j} &\sim \text{Dir}(\beta) \quad \forall\, j = 1, ..., n \\ \mathbf{A} &\sim \text{Pois}(\mathbf{WH}), \end{aligned} \tag{1}$$

that is, $\mathbf{A}_{i,j}$ is sampled from a Poisson distribution with mean given by a low-rank non-negative matrix product of Dirichlet distributions. The use of a Poisson distribution is natural because $\mathbf{A}$ represents an integer count of data with parameters taken from the matrix product $\mathbf{WH}$. We estimate these matrices by maximizing the LDA model likelihood for fixed hyper-parameters α and β, explained shortly. $\mathbf{W}$ contains t columns of length m and unit sum. These columns tell us the probability of words appearing in a tweet about that column's topic. For example, the first column may have highest probability words "soccer", "goal", and "player" while the second column may have highest probability words "rain", "wind", and "sunny". This indicates that there exist two topics respectively about soccer and weather. $\mathbf{H}$, on the other hand, contains the topic distributions for each tweet in the columns. This indicates which tweets "belong" to each topic while allowing for the possibility that a tweet could belong to more than one topic. The fundamental assumption of LDA is the existence of this low-dimensional representation $\mathbf{WH}$ of the data matrix $\mathbf{A}$. LDA is strictly a method for understanding the text content of tweets. Location and time information, though available, is not used when extracting topics. LDA performs well if highly probable words in each column of $\mathbf{W}$ make sense upon examination of their meaning. Incorrect choice of t or challenging data can result in seemingly random words being significant in each topic.

The hyper-parameters α and β make the problem easier to solve. Without the Dirichlet priors on the factor matrices, zeros can appear which makes the model probability exactly zero when there are non-zeros in the corresponding entries of $\mathbf{A}$. To avoid this occurring the Dirichlet priors are introduced. Care must be taken in choosing priors for α and β, however. Values for these hyper-parameters that are too large will force unnecessary mixing to take place. That is, if α is too large not only will the topics be prevented from having zero probabilities but they will be forced towards the uniform distribution in which all words contribute equally to each topic. The same takes place for large values of β which, if too large, will represent the tweets as even mixtures over all topics. The choices of these values depends on the details of the data. For example, the heuristic value of $50/t$ is sometimes used for β in problems where the documents (in our case tweets) are reasonably well mixed. Values one or two orders of magnitude smaller can be used to make the behavior closer to clustering where each tweet is likely assigned to a single, rather than mixture, of topics. The same applies for α which influences word concentration. But because the distributions in the columns of $\mathbf{W}$ are much larger, the previous heuristic does not apply. Generally choosing a value for α between 10^{-4} and 10^{-2} works well.

The choice of the number of topics t determines roughly how specific topics are. This is related to the amount of information each tweet provides. When using a topic model with long documents, each perhaps composed of multiple topics, only a few hundred documents can resolve a given topic. Due to their length, $1,000$ or more tweets per topic may be necessary for each topic to be meaningful. One million tweets, though numerous, will not result in more than $1,000$ useful topics. Large t will results in more specific topics, but as t becomes too large topics begin to model noise rather than trends. Small values of t result in meaningful analysis and can be used to study more general categorizations.

To perform LDA estimation we use an expectation-maximization (EM) algorithm as proposed in Asuncion et al. (2009). EM produces a single point estimate, the maximum-likelihood estimate, of the model probability rather than a probabilistic estimation (1). Other algorithms include the originally proposed method of variational Bayes (VB) and the probabilistic Markov chain Monte Carlo technique of collapsed Gibbs sampling (GS) (Casella and George, 1992). We have compared these methods and found that EM offers a good balance between performance and quality solutions. The quality of the result obtained using VB was found to be poorer than EM, and the computational time required for GS is significantly greater than EM. With our data set of 1.4 million tweets, the EM algorithm requires on the order of a few hours versus more than a day for GS. While the quality of the GS solution is generally better than that of EM, we find that it does not warrant the increased runtime. The EM algorithm is an iterative method for finding the most probable solution to (1) or, equivalently, for finding the minimum of the negative log of the probability. After removing a term in the Poisson likelihood made constant by the sum constraints, the problem is:

$$\min_{\mathbf{WH}}\left[-\sum_{i,j}\mathbf{A}_{i,j}\log\left(\sum_{k}\mathbf{W}_{i,k}\mathbf{H}_{k,j}\right)-(\alpha-1)\sum_{i,k}\log(\mathbf{W}_{i,k})-(\beta-1)\sum_{k,j}\log(\mathbf{H}_{k,j})\right],$$

which in addition to unit column sums is subject to non-negativity. The optimization strategy is to minimize this energy by alternating descent between the two matrices. Here we demonstrate the procedure for $\mathbf{W}$ and note that the process is very similar for $\mathbf{H}$. Let c_k be, for each k, a Lagrange multiplier for each column of $\mathbf{W}$ associated with the sum constraint. A single $\mathbf{W}$ update solves:

$$\min_{\mathbf{W}}\left[-\sum_{i,j}\mathbf{A}_{i,j}\log\left(\sum_{k}\mathbf{W}_{i,k}\mathbf{H}_{k,j}\right)-(\alpha-1)\sum_{i,k}\log(\mathbf{W}_{i,k})+\sum_{k}c_k\left(\sum_{i}\mathbf{W}_{i,k}-1\right)\right],$$

which has the optimality condition, where all matrix division is element-wise, given by

$$-\left(\frac{\mathbf{A}}{(\mathbf{WH})}\right)\mathbf{H}^T=\frac{(\alpha-1)}{\mathbf{W}}-\mathbf{C}.$$

The matrix $\mathbf{C}$ contains the multipliers c_k and is of the same shape as $\mathbf{W}$ with $\mathbf{C}_{i,k}=c_k$. Since $\mathbf{W}$ does not contain zeros it can be multiplied through to produce the final update for the matrix $\mathbf{W}$ with p denoting the iteration:

$$\left(\left(\frac{\mathbf{A}}{(\mathbf{W}_p\mathbf{H}_p)}\right)\mathbf{H}_p^T\right)\circ\mathbf{W}_p+(\alpha-1)=\mathbf{C}\circ\mathbf{W}_{p+1}.$$

Here $\circ$ is the Hadamard, or entry-wise, matrix product. Provided $\alpha>1$ the non-negativity constraint will be enforced naturally by the Dirichlet prior that prevents zero and negative matrix entries. The need to increase the hyperparameters by one is a known issue (Asuncion et al., 2009) resulting from under-estimation of uncertainty in the problem. In Asuncion et al. (2009) the authors demonstrate that shifting the hyper-parameters by one achieves similar performance to fully Bayesian models. Following a nearly identical derivation for $\mathbf{H}$, a single EM iteration is given by:

$$X_{i,j,k} \leftarrow (\mathbf{W}_{i,k}\mathbf{H}_{k,j}\mathbf{A}_{i,j}) \, / \left(\sum_k \mathbf{W}_{i,k}\mathbf{H}_{k,j}\right)$$

$$\mathbf{W}_{i,k} \leftarrow \frac{1}{c_k}\left(\alpha + \sum_j X_{i,j,k}\right)$$

$$\mathbf{H}_{k,j} \leftarrow \frac{1}{d_j}\left(\beta + \sum_i X_{i,j,k}\right).$$

These three steps are iterated until convergence. The values of c_k and d_j are chosen such that the columns of $\mathbf{W}$ and $\mathbf{H}$ have unit sum at each iteration. Although X is very large, in practice it is sparse due to data sparsity. We note that these computations can be parallelized which makes the final algorithm computationally fast while keeping memory requirements on the order of $\mathcal{O}(mt + tn + s)$ where s is the number of non-zero entries of A.

Topic Statistics

Topics discovered by LDA represent specific information but with varied characteristics over time and space. A topic may be very localized or broadly distributed in time or, likewise, some topics may correlate with specific locations in the city while others may not. A topic about a festival, for example, would be localized in time and space while commentary on the weather would be all year and in all locations. In this section we propose metrics to quantify these characteristics. Foremost, assign each tweet to the highest probability topic using the column-wise maximum values of $\mathbf{H}$. With these tweet-topic assignments, each topic obtains a distribution in space and time using tweet geolocations and timestamps. We propose the following four quantities for each topic based on this information.

Fractional L-norm

We first propose using the L^p norm of a function f on a domain Ω, defined by:

$$||f||_p = \left(\int_\Omega |f(x)|^p dx\right)^{\frac{1}{p}}. \tag{2}$$

This is commonly used with $p = 2$ resulting in the Euclidean norm. Different values of p capture different properties of f. In the limit $p \to \infty$ it can be shown that, under some additional assumptions, this norm approaches the maximum of f on Ω. As $p \to 0$ this norm approximates the area on which f is non-zero (Rudin, 1991). To apply Eq. 2 we generate a probability density function f_j^s over the city for each topic j using tweet locations. We use a simple histogram of the tweets on a 100-by-100 grid laid over the city. The quantity proposed is

$$\mathrm{LP}_s = \frac{||f_j^s||_p}{||f_j^s||_1}, \tag{3}$$

to quantify how "spread-out" tweets in each topic are over the city with $p < 1$.

The same method can be applied in time. We generate a probability density function f_j^t over time using a discrete histogram with days corresponding to bins. The same formula as LP_s can be used to compute LP_t, the corresponding ratio of norms applied to f_j^t

$$\mathrm{LP}_t = \frac{||f_j^t||_p}{||f_j^t||_1}. \tag{4}$$

In this paper we use $p = 0.8$ for the spatial LP value LP_s and $p = 0.1$ for the temporal LP value LP_t. The value of this metric is less important than the relative value between topics. After extracting 300 topics from our corpus we compute 600 LP values, one spatial and one temporal for each topic, residing in the range $[0, 1]$. A small LP_s describes a concentration near a few locations in space while large values indicate dispersed activity. Similarly, LP_t describes the degree to which tweets happen at concentrated times during the year versus uniformly throughout.

Mean Squared Distance

The second metric we propose is the mean squared distance of a point cloud. Consider tweets for a single topic with locations given by (x_i, y_i). The original latitude/longitude pairs are normalized so that $x_i, y_i \in [0, 1]$ for the corpus. Suppose that $i = 1, 2, ..., K$, meaning we have K tweets in this topic. The mean squared distance is:

$$\mathrm{MSD}_s = \frac{1}{K^2} \sum_{i=1}^{K} \sum_{j=1}^{K} \left((x_i - x_j)^2 + (y_i - y_j)^2 \right), \tag{5}$$

which is the expectation of the squared distance between any two tweets in the topic. Similar to the previous LP metric, we do not consider the value of this quantity but rather the relative value between all topics learned from the same corpus. Small values indicate that tweets reside close together while large values indicate that tweets are spread out. This metric occurs in physics as a “spring” potential. The metric is the ‘Energy’ associated with linking all tweet locations pairwise by springs.

Using the temporal profile of the topics, we propose MSD_t that is computed from the times of the tweets. To obtain values in $[0, 1]$, we normalize tweet times t_i and compute:

$$\mathrm{MSD}_t = \frac{1}{K^2} \sum_{i=1}^{K} \sum_{j=1}^{K} (t_i - t_j)^2, \tag{6}$$

which is the expectation of the squared time between two tweets in the topic.

Distribution Modality and Concentration

As illustrated by Fig. 1, the MSD metric quantifies spatial and temporal distributions fundamentally differently from the LP metric. When tweets are concentrated at a single discrete location, both the LP_s and MSD_s metrics will remain small. When tweets are dispersed around a single location, the LP_s will be relatively large while the MSD_s will remain small. By contrast, if tweets are concentrated at multiple but localized places in the city, MSD_s is large while LP_s remains small. If tweets are dispersed around many locations, both LP_s and MSD_s will be large.

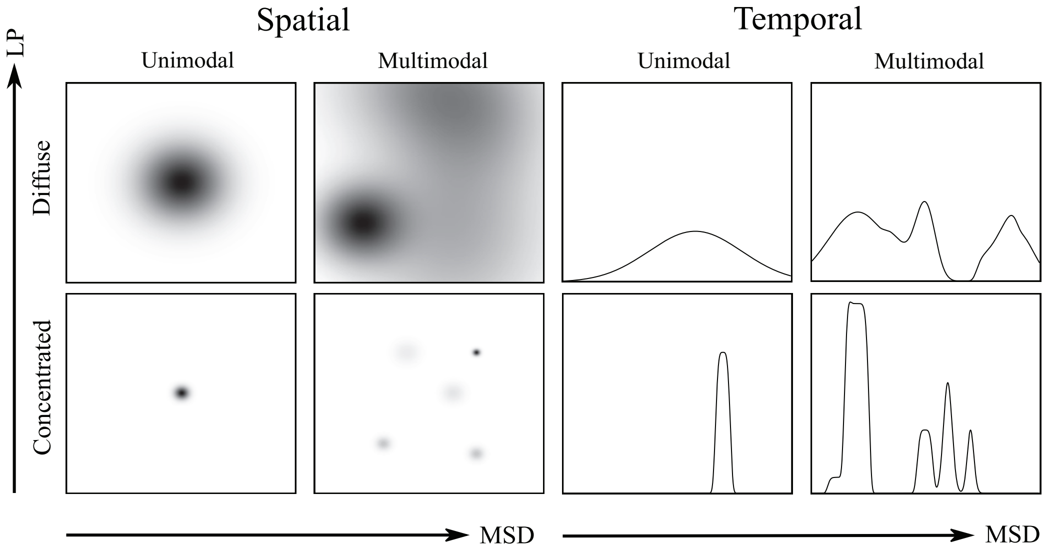

Figure 1. The conceptual relationship between the LP and MSD metrics for spatial and temporal distributions.

Similar reasoning applies for temporal distributions but in one dimension. MSD_t captures the unimodality of the temporal distribution while LP_t captures unimodality and multimodality versus uniformity. Note that the extrema of the LP metric when $p < 1$ are the Dirac delta and uniform distributions, representing maximum concentration and maximum dispersion, respectively.

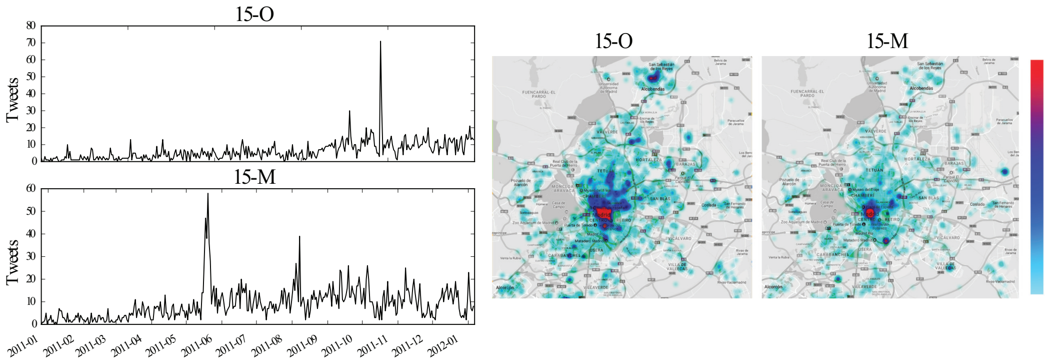

Figure 2. Protest topic histograms. Shown are histograms for topics associated with the protests "15-M" and "15-O".

Results

Our case study involves the subset of all geotagged tweets in the Madrid city area occurring from 26 December, 2010 to 6 January, 2012. The 1.4 million geotagged tweets are 3% of all tweets in the Madrid area during that period. The data contain considerable "noise" –language used on twitter frequently includes slang, hyperlinks, references to other Twitter users, and "hashtags". To address some of these issues we perform the following preprocessing routines prior to topic modelling. First, we replace characters that are not classified as alphanumeric in the Unicode character properties database and convert characters into lowercase. The data are then tokenized by separating tweets on spaces and accumulated to form $\mathbf{A}$. For our Madrid corpus, $\mathbf{A}$ contains $117,904$ rows (words) and $1,399,755$ columns (tweets). $\mathbf{A}$ contains the number of times each dictionary word appears in each document and is precisely the input to the LDA. We use the parameters $\alpha = \beta = 10^{-3}$ and extract $t = 300$ topics. We extract $t = 300$ topics from the corpus. This was chosen by examining the results of LDA up to $k = 500$ at which point the topics, on examination, became less interpretablele.

Topic Interpretation

The primary tool available for evaluating the quality of a topic model's results are the most probable words associated with each topic. Each topic produced by LDA is associated with a probability distribution over the word dictionary. Reasonable labels can be assigned to topics based on these most probable words. Consider the top words for the sample topics shown in Tab. 1. The titles for each column represent our subjective classification of each topic upon examination. The first column shows top words for a topic associated with "FITUR", an international tourism fair in Spain. The "15-M" topic is associated with a day of protest that was held on the 15^{th} of May in a location of the city known as "Puerta del Sol". The third clear topic is associated with the Madrid Barajas International Airport. The final three topics in Tab. 1 are different languages spoken in the city that, due to disjoint vocabulary, occupy disjoint topics.

In some cases, a small number of users may be responsible for topics being found. To investigate this, we examined the number of users contributing five or more tweets to each topic. 285 topics had 30 or

Table 1. Top words. Each topic is described by a probability distribution over the dictionary. Shown are the most probable words for six example topics extracted from our Madrid dataset. The titles are our interpretation.

FITUR	15-M	Airport	English	French	Portuguese
prensa	del	4	in	l	amor
internacional	sol	barajas	and	él	sueño
evento	sgae	mad	for	des	é
orgullo	campeonato	aeropuerto	day	et	eu
turismo	apertura	terminal	thanks	une	em
rueda	miau	t4	more	davidperez	não
francia	suchil	t1	us	à	um
aniversario	ancha	iberia	have	pas	pra
revista	selva	t2	nice	est	pro
fitur	carnes	airport	life	je	mais
marcatv	samurai	gate	last	ganitas	nose

more such users with a median value of 70 for all topics. For example, the "French" and "Portuguese" topics had 84 and 201 such users, respectively. Only three topics had less than 10. On examination, these are the result of automated Twitter accounts that produce significant enough quantities of tweets for LDA to dedicate a topic. While detection strategies may be used to filter out these users before the topic model, their presence does not impact the accuracy of other topics or any further analysis.

When examining the topics produced by the LDA model, three properties arise that are useful when discussing the results. First, some topics are *repetitive* in time with consistent activity occurring throughout the year. Examples of repetitive topics include sports events and leisure activities. Second, some topics are *situational*. Such topics correspond to specific events taking place in the city and hence are related to the physical situation in which they arise. Examples of situational topics are protests, festivals, and conventions. *Non-situational* topics are disconnected from the physical city. Examples of non-situational topics are global events or celebrity gossip. The final categorization we propose is *language* topics. Examples encountered in our corpus are French, Italian, Portuguese, Tagalog, and Indonesian.

Spatial and Temporal Patterns

In addition to the top words within each topic from $\mathbf{W}$, the distribution over topics for each tweet is available via $\mathbf{H}$. By taking the most probable topic for each tweet, we approximately cluster the tweets and, using the locations and times of the tweets, form temporal and spatial histograms for each topic. To understand these distributions, we propose the LP and MSD metrics in space and time. Fig. 3 presents metrics for all topics with points of interest labeled. On the left are spatial metrics LP_s and MSD_s while on the right are the temporal metrics LP_t and MSD_t.

Applying a topic model to a large data collection can produce a large number of topics. 300 topics requires significant time to sift through manually, and topic counts often can number in the thousands. The proposed metrics allow a user unfamiliar with Madrid to quickly arrive at topics related to the underlying city.

The "Airport" topic is spatially concentrated as captured by the low values of LP_s and MSD_s. This is intuitively reasonable given that the Madrid airport is localized in space. Close inspection of the

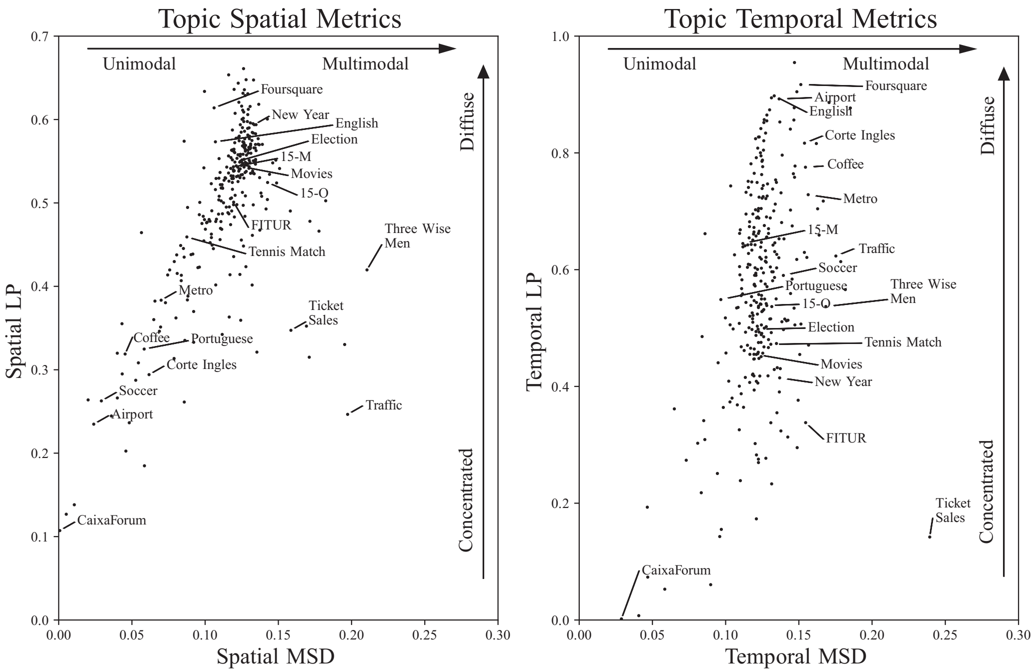

Figure 3. Metrics for all topics. Shown are the values of the spatial (left) and temporal (right) metrics proposed to study the topics learned from the corpus.

spatial density map of tweets confirms this (Fig. 4). In contrast, tweets within the “Three Wise Men” topic show LP_s and MSD_s metrics of a different nature than the “Airport” topic (Fig. 3). This topic is related to Christian religious celebrations that typically take place early in the new year. The “Three Wise Men” topic is not concentrated at a single point like the “airport” topic (Fig. 4) resulting in large MSD_s. For the “Foursquare” topic both spatial metrics are large. Foursquare is a social media program for smartphones that lets users “check-in” to locations such as restaurants and museums and then broadcast this information to other users using automatically generated tweets. The large MSD_s and LP_s values indicate that no particular location or locations are associated with this topic (Fig. 3).

The temporal metrics are similarly useful for characterizing topics. Fig. 5 presents examples of temporal histograms for eight topics roughly grouped based on appearance. The top panel contains two topics with very high LP_t indicating no particular association with any specific time in the year. The second panel shows temporal histograms for situational topics associated with two holidays. Both topics have low LP_t but, due to activity at both ends of the year from annual recurrence, the MSD_t values are neither large nor small. The third panel contains two topics with spikes in activity arising at many points in the year. This results in the LP_t being slightly higher than for topics associated with very sparse events like holidays, but still lower than for e.g. the “Airport” topic (see 3). The final two topics in the bottom panel are oddities and result in extreme values of MSD_t and LP_t. These result from the patterns

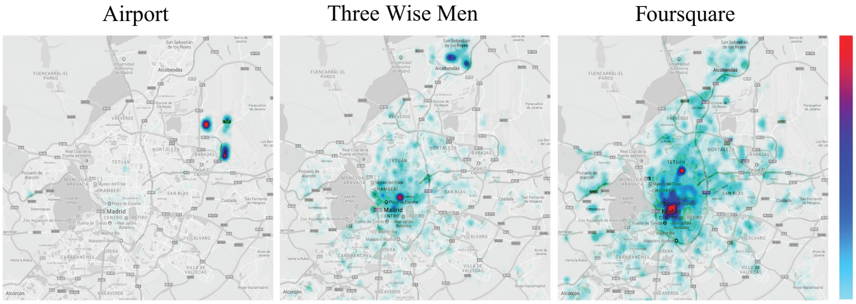

Figure 4. Example spatial histograms via Google's mapping API. These histograms demonstrate the characteristics captured by our metrics in Fig. 3: airport activity (small LP$_s$, small MSD$_s$), the Three Wise Men festival (small LP$_s$, large MSD$_s$), and check-ins to the Foursquare service (large LP$_s$, large MSD$_s$).

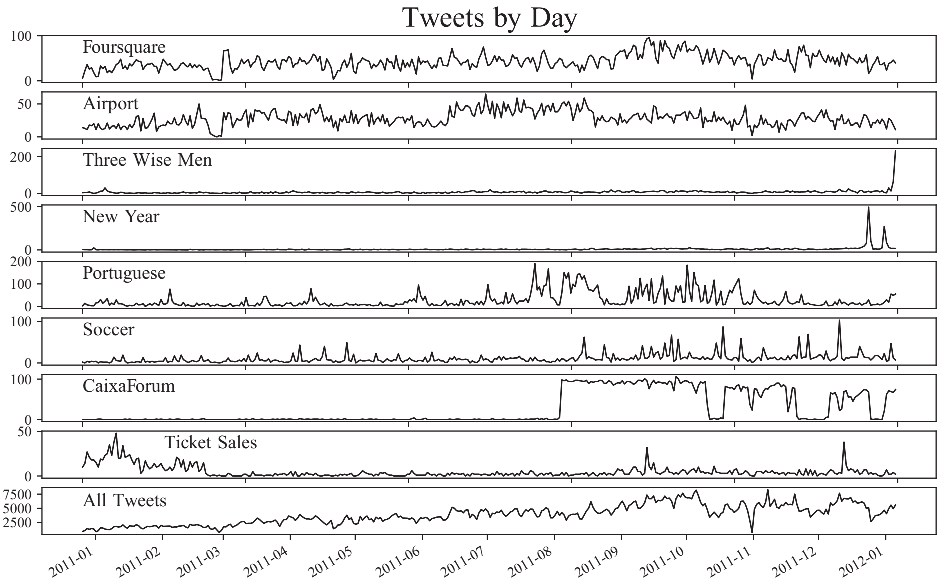

Figure 5. Types of temporal histograms. Different topics are characterized by different metric values that indicate the type of temporal activity. Shown are a few examples of background topics (top two rows), singular events in the year (rows 3-4), event topics with many activity spikes in the year (rows 5-6), and outliers arising from automated tweeting (rows 7-8). Also shown is the background for all tweets (bottom row). The metric values help to understand these distributions.

of automated Twitter accounts, mentioned previously, that are enabled or disabled for different portions of the year.

The characterization of repetitive and situational topics is therefore, approximately, associated with LP_t. A high LP_t indicates the repetitive property while a low value indicates the situational property. The specification of language topics, however, does not appear to have a clear quantitative indicator using the metrics. These metrics therefore capture significant information for determining Twitter activity related to the underlying city. This organization of the topics is invaluable for large data sets with many topics.

Unique Events

Unique events in the city can be separated out using the LP_t value. A high value, as in the case of "Airport" and "English", are not specifically associated with times in the year. Low LP_t indicates unique events that took place. Additionally, MSD_t indicates roughly the temporal modality. Looking at topics with a small MSD_t value reduces the topic space to only those associated with single events in time. For example, in Fig. 3 a user can quickly arrive at topics such as a tennis match or significant local and regional elections. Important topics of this character relate to protests on the 15th of May and October 2011, respectively known as the "15-M" and "15-O" protests (see 2). The 15-M protest was one of the most important events in the recent history of Madrid (and Spanish) politics. 15-M was a large anti-austerity protest in which activists occupied the Puerta del Sol plaza in central Madrid. Due to the magnitude of this event throughout 2011, the LP_t metric is higher than for other unique events. Further refinement can be performed with the spatial metrics. For example, finding city-wide events of significance is a matter of further looking into topics with high LP_s. This includes topics such as "Election" to which people across the city contribute. In contrast, "Tennis Match" is much more localized in space as revealed by the spatial metrics. By carefully considering spatial and temporal metrics, events of different character are quickly found.

Repetitive Topics

Repetitive topic consisting of background activity in a city throughout a year can be found using a similar metric analysis. High LP_t is associated with topics such as "Airport", "Corte Inglés", "Coffee", and "Metro". The "Corte Inglés" topic results from shopping activity at a famous chain of stores in Madrid. Focusing in on this subset of topics results in only those that correspond to continuous activity through the year. Further refinement can be performed using space. For example, repetitive topics with little spatial association will have high LP_s such as the "Foursquare" topic. Alternatively, topics with low LP_s correspond to specific areas of the city. The unimodal "Airport" topic, for example, has a low MSD_s because it is associated with one location. The multimodal "Coffee" and "Metro" topics are associated with multiple locations concentrated around the city center with the latter further spread out. These subtle differences in distribution are captured by the spatial metrics directly, and a user can find them without first studying the top words for all 300 topics.

Situational Topics

Situational topics are also found using the proposed metrics. Examples are "FITUR", "New Year", and "Three Wise Men". The first is related to an annual tourism fair and the last is the Spanish tradition of the

"Reyes Magos"[‡] with top words such as "magos", "reyes", "caramelo", and "cabalgata"[§]. Like unique events, the LP$_t$ metric indicates reasonably well which topics are situational – the three discussed here all have low values. The MSD$_t$ does not appear to be particularly informative for these topics, a fact that might result from the periodic nature of years – the MSD$_t$ metric considers the first and last day of the year very far apart and hence "New Year" is of a high value. Situational topics can have different spatial properties. The "Three Wise Men" topic is concentrated at two locations in the city, "New Year" is active across the city and "FITUR" is concentrated at the event and nearby shopping areas. These differences are made clear with the proposed spatial metric values.

Language Topics

The final category of topics proposed are language topics. After Spanish, English is the second most prevalent language in our data and is associated with multiple topics. The locations of these topics' tweets often show a distribution all over the city of Madrid with perhaps a slight concentration near downtown, believed to be a result of tourism. Immigrant populations in Madrid are concentrated in the center and southwest of the city (Bosch et al., 2011). Other languages found occur in single topics, including French, Portuguese, Tagalog, Italian, German, Indonesian, Catalan, and Dutch. The appearance of Tagalog and Indonesian topics is especially curious given that they are the only languages from non-European locations. The temporal histograms, and therefore also metrics, depend on the presence of speakers visiting the city. Ultimately the properties of each language topic vary depending on the presence and distribution of speakers in the city and, therefore, the proposed metrics do not distinguish such topics readily.

Discussion and Conclusion

We have proposed an efficient pipeline for using geotagged Twitter data to build a picture of activity, events, and behavior within a city. This pipeline has two major components. The first component is pre-processing and topic extraction, using an expectation maximization algorithm for latent Dirichlet allocation. This takes several hours to run on our data set of 1.4 million tweets from Madrid. The second step can be performed in real time and involves the calculation of spatial and temporal metrics based on the fractional L^p norm and a mean squared tweet distance applied to tweets within topics. With this information, we are able to separate background activity such as coffee shops from actvity unrelated to the city and special events. Most importantly, major events in Madrid such as elections, and the 15-M and 15-O protests can be quickly identified and followed. Different types of topic activity can be identified through the use of the proposed metrics to quickly organize topics for study.

We provide source code for all steps of the computational methods along with a discussion of what is the expected outcome. Notably much more could be done in future work. For example, one could perform more precise temporal statistical analysis of the data. In other work (Lai et al., 2014), there has been some study of self-excitation in repeat activities. At the same time, we expect that events localized in time will have different statistics, while the events that are associated with repeat scheduled activities, such

[‡] "Three Wise Men" in English.

[§] The term "reyes magos" is translated as "wise men". However, the direct translation for the separate words would be "magic" (magos) and "kings" (reyes). The remaining words are "candy" (caramelo) and "parade" (cabalgata).

as soccer games, will have yet another temporal signature. The proposed spatial and temporal metrics capture important properties of the distributions of tweets, but certainly more precise statistical models may be used to extract further information. Another research question is one of prediction - whether future events could be predicted from prior data especially given the complexity of spatial and temporal patterns amongst the different topics. We hope that our software and methods will be useful for future work in this area.

Finally, we also believe this work holds implications for an intervention model to promote or prevent activities associated with certain topics. Our metric characterization of topics provides guidance on how interventions could be structured in space and time, although they do not tell you what interventions to adopt. In general, topics that are concentrated at unimodal locations in space can be well-targeted by placed-based interventions (Weisburd, 2008). Place-based interventions are still possible if spatial distributions are multimodal, but concentrated. The limit here is on availability of resources to distribute over modes. It becomes more problematic if topic distributions are dispersed. At the extreme, a topic may be so dispersed in space that there is no feasible place-based intervention. Similar arguments hold for the modality and concentration of topics in time.

References

Asuncion, A., Welling, M., Smyth, P., and Teh, Y. W. (2009). On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 27–34. AUAI Press.

Asur, S. and Huberman, B. (2010). Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 492–499. IEEE.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Bosch, M., Carnero, M., and Farré, L. (2011). Rental housing discrimination and the persistence of ethnic enclaves. *Institute for the Study of Labor*, Discussion Paper 5583.

Casella, G. and George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174.

Compton, R., Jurgens, D., and Allen, D. (2014). Geotagging one hundred million twitter accounts with total variation minimization. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 393–401. IEEE.

Ding, C., Li, T., and Peng, W. (2008). On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Computational Statistics & Data Analysis*, 52(8):3913–3927.

Farrahi, K. and Gatica-Perez, D. (2011). Discovering routines from large-scale human locations using probabilistic topic models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(1):3.

Frias-Martinez, V. and Frias-Martinez, E. (2014). Spectral clustering for sensing urban land use using twitter activity. *Engineering Applications of Artificial Intelligence*, 35:237–245.

Gerber, M. S. (2014). Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61:115–125.

Han, B., Cook, P., and Baldwin, T. (2014). Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, pages 451–500.

Heverin, T. and Zach, L. (2010). Twitter for city police department information sharing. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–7.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc.

Hong, L., Ahmed, A., Gurumurthy, S., Smola, A. J., and Tsioutsiouliklis, K. (2012). Discovering geographical topics in the twitter stream. In *Proceedings of the 21st international conference on World Wide Web*, pages 769–778. ACM.

Hong, L. and Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88. ACM.

Hu, B., Jamali, M., and Ester, M. (2013). Spatio-temporal topic modeling in mobile social media for location recommendation. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 1073–1078. IEEE.

Kling, F. and Pozdnoukhov, A. (2012). When a city tells a story: urban topic analysis. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, pages 482–485. ACM.

Laflin, P., Mantzaris, A. V., Ainley, F., Otley, A., Grindrod, P., and Higham, D. J. (2012). Dynamic targeting in an online social medium. *Social Informatics*, pages 82–95.

Lai, E., Moyer, D., Yuan, B., Fox, E., Hunter, B., Bertozzi, A. L., and Brantingham, P. J. (2014). Topic time series analysis of microblogs. UCLA CAM Report 14-76; https://www.math.ucla.edu/applied/cam.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Ramage, D., Dumais, S. T., and Liebling, D. J. (2010). Characterizing microblogs with topic models. *ICWSM*, 10:1–1.

Ritter, A., Cherry, C., and Dolan, B. (2010). Unsupervised modeling of twitter conversations. *Proceedings of HLT '10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180.

Rudin, W. (1991). *Functional Analysis*. McGraw-Hill Science/Engineering/Math.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM.

Schwartz, R., Naaman, M., and Matni, Z. (2013). Making sense of cities using social media: Requirements for hyper-local data aggregation tools. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

Suh, B., Hong, L., Pirolli, P., and Chi, E. H. (2010). Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Social computing (socialcom), 2010 ieee second international conference on*, pages 177–184. IEEE.

Weisburd, D. (2008). Place-based policing. *Ideas in American Policing*, 9:16.

Weng, J. and Lee, B.-S. (2011). Event detection in twitter. *ICWSM*, 11:401–408.

Zhao, W. X., Jiang, J., He, J., Song, Y., Achananuparp, P., Lim, E.-P., and Li, X. (2011). Topical keyphrase extraction from twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 379–388. Association for Computational Linguistics.

Acknowledgements

The raw data used in this paper was purchased from trendsmap.com for research purposes. This work was supported by the Department of Defense, ONR grant N00014-16-1-2119, and NSF grant DMS-1417674.

Author Biographies

Travis Meyer is a graduate student in the Department of Mathematics at the University of California, Los Angeles. His research is focused on energy-based models for signal processing and data science with a particular focus on matrix factorization models and general learning with variational models. In addition to his research, he has served in a variety of roles academically instructing machine learning courses and supervising undergraduate research. He received a B.S. in Applied Mathematics and B.A. in Physics from the University of California, Los Angeles in 2011.

Daniel Balagué is an adjunct assistant professor in the Department of Mathematics, Applied Mathematics and Statistics at Case Western Reserve University (CWRU). He obtained his PhD in Mathematics at Universitat Autonoma de Barcelona. He completed a postdoc at North Carolina State University and was Assistant Adjunct Professor in the Program in Computing in the Department of Mathematics at the University of California, Los Angeles. Currently, he works for the Research Computing and Cyberinfrastructure (RCCI) group at Case Western. His research interests include partial differential equations, mathematical physics, big data, parallel programming, code optimization and software development.

Miguel Camacho Collados is a Spanish Police Inspector, current Head of the National Police Statistics and Analysis Unit, with experience in criminal investigation and human resources management. Miguel obtained his Ph.D. in Mathematics and Statistics (Summa cum laude) from the University of Granada. He is a Fulbright Alumni and continues to direct research projects involving academics and the Spanish National Police Corps. His research interests concern the use of data science as a means to prevent and combat crime.

Hao Li received his degree of Bachelor of Science in the Mathematics of Computation and Master of Arts in Applied Mathematics at the University of California, Los Angeles in 2016. He is currently a first-year graduate student in the Department of Mathematics at the University of California, Los Angeles, working towards his Ph.D degree. His research interests lie in numerical optimization with its application in data analysis and machine learning, specifically in the domain of topic modeling and network analysis.

Katie Khuu is a senior Computer Science major at the University of California, Irvine. She has conducted research in Bayesian models of cognition and data analytics. Katie is a scholar of the University of California Leadership Excellence Through Advanced Degrees Program and a fellow of the Undergraduate Research Opportunities Program at UCI. She won Undergraduate Research Award from the University of California Education Abroad Program and a Best Poster Award from the IC Academic Research Program Symposium. Katie will be joining Expedia, Inc. as a software development engineer after graduation.

Jeff Brantingham is Professor of Anthropology at the University of California Los Angeles. Jeff obtained his B.A. in Anthropology from the University of British Columbia and his M.A. and Ph.D. in Anthropology from the University of Arizona. His research interests lie in the study of human behavior in complex environments, including offender mobility, offender target selection and the organization of criminal street gangs. He has published more than 70 academic journal articles.

Andrea Bertozzi is Professor of Mathematics and the Betsy Wood Knapp Chair for Innovation and Creativity at UCLA. She is a Fellow of the American Academy of Arts and Sciences and a Fellow of the Society for Industrial and Applied Mathematics, the American Mathematical Society, and the American Physical Society. She received three degrees in Mathematics from Princeton University in 1987 (A. B., Summa cum laude), 1988 (M. A.), and 1991 (Ph.

D.). Her research focuses on nonlinear partial differential equations with applications to image processing, crime modeling, and swarming/cooperative dynamics.