IMA Journal of Applied Mathematics (2016) **81**, 409–431
doi:10.1093/imamat/hxw025
Advance Access publication on 13 July 2016

Topic time series analysis of microblogs

Eric L. Lai
Department of Mathematics, UCI, Irvine, CA 92697, USA
Daniel Moyer
Department of Computer Science, University of Southern California, Los Angeles, CA 90033, USA and Department of Mathematics, UCLA, Los Angeles, CA 90095, USA
Baichuan Yuan
Department of Mathematics, Zhejiang University, Hangzhou 310027, China and Department of Mathematics, UCLA, Los Angeles, CA 90095, USA
Eric Fox
Department of Statistics, UCLA, Los Angeles, CA 90095, USA
Blake Hunter
Mathematical Sciences, Claremont Mckenna College, Claremont, CA 91711, USA and Department of Mathematics, UCLA, Los Angeles, CA 90095, USA
Andrea L. Bertozzi*
Department of Mathematics, UCLA, Los Angeles, CA 90095, USA
*Corresponding author: bertozzi@math.ucla.edu

and

P. Jeffrey Brantingham
Department of Anthropology, UCLA, Los Angeles, CA 90095, USA

[Received on 24 April 2016]

Social media data tend to cluster around events and themes. Local newsworthy events, sports team victories or defeats, abnormal weather patterns and globally trending topics all influence the content of online discussion. The automated discovery of these underlying themes from corpora of text is of interest to numerous academic fields as well as to law enforcement organizations and commercial users. One useful class of tools to deal with such problems are topic models, which attempt to recover latent groups of word associations from the text. However, it is clear that these topics may also exhibit patterns in both time and space. The recovery of such patterns complements the analysis of the text itself and in many cases provides additional context. In this work we describe two methods for mining interesting spatio-temporal dynamics and relations among topics, one that compares the topic distributions as histograms in space and time and another that models topics over time as temporal or spatio-temporal Hawkes process with exponential trigger functions. Both methods may be used to discover topics with abnormal distributions in space and time. The second method also allows for self-exciting topics and can recover intertopic relationships (excitation or inhibition) in both time and space. We apply these methods to a geo-tagged Twitter dataset and provide analysis and discussion of the results.

Keywords: mining complex datasets; spatial and temporal analysis; topic modeling; cluster analysis.

1. Introduction

It is apparent that microblogs such as Twitter are composed of a vast number of diverse topics. Unfiltered samples from the Twitter ‘firehose’ often contain tweets on wide variety of topics such as local politics, sporting events, daily activities, weather, local crime and organized public demonstrations. The summarization and analysis of these data is of interest to social scientists, commercial groups, law enforcement and government agencies among others.

However, the extraction of semantic information from raw text is a non-trivial task. A large amount of literature has been devoted to modelling and extracting latent themes from both Twitter and large text corpora in general. Known as topic models, these methods use latent word associations (referred to as topics) to capture the underlying themes in the documents (i.e. the Tweets). In practice many practitioners use a very large number of topics due to the diversity of the text. While originally intended to summarize latent themes in the data, the topics may be so numerous that they themselves may require automated analysis.

At the same time, we are often able to recover more information from the media source than just the text content. Microblog data often include metadata such as posting time and location, allowing us to produce distributions of documents over physical space and time. In the context of a spatio-temporal process, some topics are observed in Tweets purely at random (topics associated with teenage romance perhaps) or on a periodic basis with spatial clusters (topics about rush hour traffic, local weather or major holidays). Still others exhibit patterns quite different from baseline Twitter usage. Natural disasters, one-time fads and large events (including mass civil disturbances) can be expected to produce anomalous Twitter content.

It is useful then to produce automated topic analysis methods focused around identifying spatio-temporal patterns. Topics with temporal or spatial distributions that are anomalous with respect to the background rate of document occurrence may be of further interest to analysts and may be indicative of a corresponding real world event. Furthermore, given a specific location and/or time, it is helpful to be able to find associated topics (and thereby documents).

Topics may also exhibit temporal or spatio-temporal couplings. Social events may trigger further events, sports team victories or defeats may lead to the discussion of the future of a player or coach’s employment or a controversial post may trigger an explosion of heated responses. In terms of topics and Tweets, the observation of some Tweets from a topic may precede the observation of Tweets from another related topic with some regularity. In a predictive sense, the observation of Tweets from some topics may contain information about the incidence rate of Tweets from another topic (Ding *et al.*, 2013; Ver Steeg & Galstyan, 2012). For example, if we observe a number of observations in a bad weather topic, we might expect to see a number of observations in the traffic topic. On a larger scale, the pairwise coupling of topics is indicative of a possible latent network structure.

In this work we give two methods for the automated mining of temporally and spatially anomalous topics generated by non-negative matrix factorization (NMF). One method is based on the earthmover’s distance (EMD) and provides a distance measure of a topic from a background distribution. The other is based on the Hawkes process, which is a self-exciting point process model. The second provides estimates of latent network structures and has associated goodness-of-fit measures. To validate these methods we process 500,000 geolocalized Twitter messages from the Los Angeles area over a 10-month period. The Tweets are timestamped and geo-tagged (geographical location information from the user attached to the Tweet).

2. Previous work

Our methods build upon recent literature concerning the spatio-temporal analysis of human activity patterns, topic modelling, anomaly detection and self-exciting point processes.

2.1. *Spatio-temporal human activity*

It is well known that human activity is not uniformly distributed in space or time. Particular activity types tend to cluster in local spatial regions, while the frequencies of those behaviours also tend to cluster in time. The clustered, bursty nature of human behaviour has huge implications for the organization and function of urban systems. Our own previous work has concentrated on the spatio-temporal dynamics of crime which, like other aspects of human behaviour, forms dynamic spatio-temporal hotspots (Lewis & Mohler, 2011; Mohler, 2014; Woodworth *et al.*, 2014).

Outside of this, mobile device usage has been shown to also have a clustered nature (Gonzalez *et al.*, 2008), following human behaviour. The rise of social media and mobile data allows similar analyses to be taken a step further. One prime example is the use of Foursquare data to make inferences about user activities, geographic regions and local events (Noulas *et al.*, 2011).

2.2. *Microblogs and related topic models*

Twitter as a source of data for academic study has been in use since approximately 2007 (Java *et al.*, 2007), when it was treated as a social network. Since then, it has been a popular topic of study, so much that there are papers about people writing about Twitter (Williams *et al.*, 2013). A growing proportion of studies look principally at Twitter content; it has been suggested that Twitter, while presenting a social network and an information diffusion network, may be closer to a media distribution site, where the media is user produced (Kwak *et al.*, 2010). Analysis of the text content includes both general models as well as Twitter-specific models (Hong & Davison, 2010; Zhao *et al.*, 2011). Grindrod (2014) looks at a dynamic random walk time series model for event-driven spikes in Twitter data and outlines many of the current state-of-the-art approaches in the area.

Several previous works have introduced geospatial or time-dependent topic models. In particular Cataldi *et al.* (2010) introduce a time-dependent topic model. Similarly, both Yin *et al.* (2011) and Hong *et al.* (2012) provide variants of geospatial topic models. In general these approaches differ from our method in that they directly influence the chosen topics by their respective domain (time or space). Though these are important contributions, the topics produced from these models may not accurately describe the full corpus of text.

Along a different line is the much more recent exploration of topics on hidden information diffusion graphs by He *et al.* (2015). This work also uses the Hawkes process but, instead of constructing separate timelines for each topic, models each user as having a separate timeline. This excellent work parallels our second method and uses a similar multivariate marked Hawkes process.

2.3. *Spatial and temporal anomaly detection*

Directly related to our first method is a work by Applegate *et al.* (2011). The authors consider only mobile phone usage data without content, applying an approximate EMD described in Shirdhonkar & Jacobs (2008) to cluster temporal patterns across multiple cyclic periods (e.g. patterns over time of day and day of the week between different users). Our work extends their approximate EMD from time-based histograms to a histograms weighted by document content both in space and time.

More related to our second method are event detection and summary methods. Twitter is known to reflect real world events and news media activity.

Similar to our work, Zhao *et al.* (2011) use a Twitter generative text model based on latent Dirichlet allocation (LDA), then match Topics between the generated Twitter model and *New York Times* articles. This provides important groundwork for investigating temporal coupling between documents, though not in a point process context.

2.4. *Point process models for social interaction*

Recently, several studies have modelled social interactions as linked Hawkes point processes. While not analysing text content, nor, in general, microblog activity, these studies employ methods that are similar to the ones explored in Sections 5 and 6.

In particular, Blundell *et al.* (2012) model reciprocity in relationships from human interaction data using linked Hawkes processes. The authors fit their model to several datasets, including selected threads of the ENRON email corpus and the Militarized Interstate Dispute corpus. Following this work, Zhou *et al.* (2013) provide a sparse, low-rank extension, using the same multidimensional Hawkes process to model information diffusion across networks. This second set fits the model to a MemeTracker dataset, a similar setting to Twitter.

A more technical review of the Hawkes process and relevant citations is given in Section 5.

3. Topic models

In order to extract latent topic variables from our text corpus, we transform our raw text data into a Bag-of-Words vector form and then apply NMF with sparse constraints. The pre-processing work, while involved and non-trivial, is not our focus nor do we introduce any innovations to the field and so is only covered briefly here.

3.1. *Pre-processing*

As found in Ramage *et al.* (2010), Godin *et al.* (2013) and Hong & Davison (2010), we apply significant pre-processing to our raw data before training our topic model. The steps here are undertaken in order: 1. We encode the text into ASCII, discarding any unicode characters. 2. We replace all double quotes with the empty string. 3. We extract all user references and all hashtags, denoted, respectively, with @ or # at the beginning of a token. 4. We attempt to remove any urls, specifically anything prefixed with 'http'. 5. We remove many non-alphanumeric characters, with the important exception of $ and @, with the latter only in the case that it is the only character in the token (the @ symbol is significant in its usage by Instagram in automatically generated Tweets). 6. We change all characters to lowercase. 7. We remove any token on our Stop Words list, including a Twitter-specific stopwords list of the 50 most common words observed in our dataset. 8. We remove any token observed less than 10 times. 9. We partition the data by month in order to reduce the number of fad-like topics observed in each data set.

After pre-processing we form an ordered vocabulary and generate term-frequency vectors from the documents. We concatenate these to form a data matrix D', where each row is a document, and each column represents a distinct word in our vocabulary. We immediately re-weight D' using the TF-IDF scheme (Salton & McGill, 1983). This re-weighted matrix we denote as D.

We denote the number of documents N, and the number of words in our vocabulary M; thus, $D \in \mathbb{R}^{N\times M}$. For this analysis $N > M$. As a matrix of frequency counts, D only has non-negative entries.

3.2. *Non-negative matrix factorization*

After forming our data matrix D, we then make the assumption that the rows of D are approximately the additive combination of K non-negative topic vectors, where $K \ll N$. This is equivalent to making the assumption that D is approximately of rank K, with the constraint that the subspace spanned by D has a set of non-negative basis vectors and all of the rows of D have non-negative coordinates in that basis.

Using this assumption, we have the following approximation $D \approx WH^T$, where W is a matrix of the coordinates of each document in the subspace of the rows of H^T. This is the basic NMF (Lee & Seung, 1999), which has the objective function $J(W,H) = ||D - WH^T||_F$. The matrix norm used here is the Frobenius norm. With a slight modification of the above objective and use of the Kullback–Leibler divergence instead of the Frobenius norm, NMF has been shown to be equivalent to Probabilistic Latent Semantic Indexing (Ding *et al.*, 2008), a forerunner of LDA.

In the recent literature, good results have been achieved using a combination of an L_1 and an L_2 regularizing term (Kim & Park, 2008b; Saha & Sindhwani, 2012). This encourages sparsity and somewhat prevents overfitting. Our specific objective is given below:

$$J(W,H) = \frac{1}{2}\|D - WH^T\|_F^2 + \alpha\|W\|_F^2 + \beta\sum_{i=1}^{n}\|H_{i,:}\|_1^2 \tag{3.1}$$

subject to the non-negative constraints on both W and H.

In this article we use an alternating least squares (ALS) active set method developed by Kim & Park (2008a), using 300 topics. ALS methods alternate between minimizing $||D - WH^T||_F$ (the sum of element-wise squared error) over W and H matrices. In this particular case each regularizer term contains only either W or H terms, so Kim and Park encode the regularizer terms into

$$J_1(H) = \left\| \begin{pmatrix} D \\ 0 \end{pmatrix} - \begin{pmatrix} W \\ \mathbf{1} \end{pmatrix} H^T \right\| \qquad J_2(W) = \left\| \begin{pmatrix} D \\ 0 \end{pmatrix} - \begin{pmatrix} H \\ I_k \end{pmatrix} W^T \right\|.$$

They then employ the usual ALS procedure, alternating between minimizing $J_1(H)$ and $J_2(W)$.

Each of the K rows of H^T may be interpreted as a topic vector and each entry of a given row as the relative frequency with which a word occurs in the topic. By sorting the entries of the row we can form ranked lists of words describing the topic. We show an example of this in Table 1. Each of the N rows of W is the encoding of a document in the topic basis. Each entry of a given row of W is the proportion of the document that is ‘taken’ from a given topic.

This topic model can efficiently handle streaming data or data that arrive in large batches over time. Given the model parameters previously learned on current data, it is a simple matrix manipulation to find the topic representation of new data as it comes in. The parameters can also be updated offline to adapt to changes in the underlying structure due to the addition of new data.

4. Earthmover’s distance

In this section we first define the EMD and briefly discuss its motivation, important properties and differences from other measures and metrics. We then discuss our usage of it and present results.

TABLE 1 *Examples of topic interpretations for select topics. On the left-hand side the corresponding topic number is provided. As expected, numerous topics have a running theme, e.g. Holidays, Events, Classes and Sports Teams. Others seem to be dominated by a single word to which the others all relate. Topics in boldface possibly exhibit spatial or temporal patterning, which the methods presented here investigate rigorously.*

Topic no.	Topic words in descending order of frequency per topic				
75	Song	Sing	Lyric	Singing	Hear
79	Hungry	Bored	Af	Lazy	Super
134	**Citadel**	**Outlets**	**Shopping**	**Commerce**	**Others**
136	Birthday	Enjoy	Happy	Beautiful	Cake
138	**Years**	**New**	**Kiss**	**Resolution**	**Eve**
154	**California**	**State**	**University**	**Angeles**	**Los**
172	Class	Math	Ugh	Spanish	Full
184	Idk	Might	Yet	Umm	Bout
188	Cute	Boyfriend	Aww	Together	Aha
192	**Win**	**Lakers**	**Straight**	**Fan**	**Clippers**
231	**Merry**	**Christmas**	**Xmas**	**Yall**	**Everybody**
227	Okay	Ahaha	Ohh	Aww	Hahahaha
237	**Commerce**	**Old**	**Store**	**Factory**	**Change**
242	Tho	Ahah	Af	Lame	Serious
25	School	High	Middle	Monday	Excited
33	Stay	Strong	Kind	Faithful	Single
40	Wait	Till	Til	Excited	Train
62	Cool	Sound	Kinda	Minute	Reply
64	Stupid	Acting	Af	-_-	Act
113	Text	Number	Sent	Message	Reply
117	Good	Sound	Luck	Feels	Mood
143	Gotta	Clean	Fight	Learn	Dawg
147	Break	Heart	Winter	Taking	Fast
176	**Food**	**Mexican**	**Bomb**	**Chinese**	**Ate**
179	Face	Sad	Ugly	Beautiful	Punch
181	Take	Nap	Shower	Breath	Seriously
211	Asleep	Fall	Fell	Falling	Half

4.1. *Definition of the EMD*

Let P and Q be discrete distributions:

$$P = \{(p_1, w_{p1}), \ldots, (p_N, w_{pN})\} \tag{4.1}$$

$$Q = \{(q_1, w_{q1}), \ldots, (q_M, w_{qM})\}, \tag{4.2}$$

$$\sum_{i=1}^{N} w_{pi} = 1 \text{ and } \sum_{i=1}^{M} w_{qi} = 1. \tag{4.3}$$

Let $d(\cdot,\cdot)$ be a metric on the set $\{p_i\}_{i=1}^{N} \cup \{q_i\}_{i=1}^{M}$ and let f_{ij} be the scalar flow from p_i to q_j with the following constraints:

$$f_{ij} \geq 0, \quad \sum_{j}^{M} f_{ij} = w_{pi}, \quad \sum_{i}^{N} f_{ij} = w_{qj}. \tag{4.4}$$

We define the EMD as

$$EMD(P,Q) = \min_{\{f_{ij}\}} \sum_{i,j} f_{ij} \cdot d(p_i, q_j), \tag{4.5}$$

as seen in Muskulus & Verduyn-Lunel (2011). More intuitively, if P and Q were piles of dirt, the EMD measure would be similar to the minimum work required to move the pile P to the pile Q. For more analytic results, the EMD is commonly extended to continuous event spaces; in this article we only use the discrete version.

EMD is a metric on distributions defined over a metric space. The metric space condition is due to the ground distance or flow property of EMD, a property which also separates it from other metrics such as total variation.

4.2. *Construction of histograms*

Once each document in the corpus has been assigned a topic encoding, we recover a empirical distribution in space and time for each topic. Here we only rigorously address a 1D histogram, but the process is easily extended to higher dimensions.

Given an connected observational window $L = [a, b]$ and a fixed number of bins B, we partition the window into B subintervals of length $h = \frac{b-a}{B}$. Each sub-interval is defined as $\ell_j = [a + h \times j, a + h \times (j + 1)]$. For a given corpus D with documents d_i, topics Z, topic encodings $c_{i,z}$ and positions $t_i \in [a, b]$, we define the distribution P_z of a given topic $z \in Z$ as the following vector (histogram):

$$p_{j,z} = \frac{\sum_{t_i \in \ell_j} c_{i,z}}{\sum_{d_i} c_{i,z}}. \tag{4.6}$$

This is readily interpreted as the binned distribution of Tweets in L, reweighed by their topic encodings and normalized so that the bins sum to one. We also define the ‘uniform’ weighting of the Tweets, which we refer to as the uniform histogram; note that this is not a uniform distribution over space or time but is the binned background rate of all Tweets (uniformly weighted).

Because the number of bins increases exponentially with the dimension of the ground distance, common algorithms for computing the exact solution to EMD scale badly. To avoid this cost, we use an approximation to the EMD originally formulated by Shirdhonkar & Jacobs (2008), which relies on the wavelet transform. This takes the computation from approximately $O(n^3)$ to $O(n)$, where n is the number of bins.

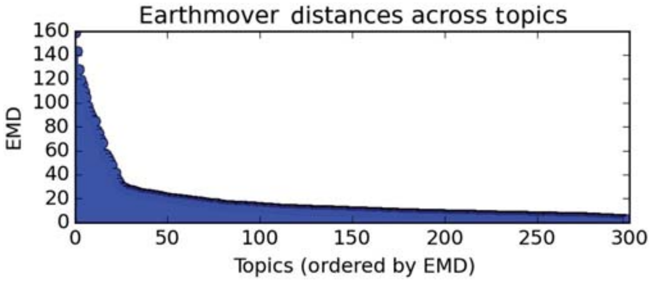


FIG. 1. EMDs for each temporal histogram.

4.3. *Application to twitter timeseries*

In the context of Twitter data, we construct topic timeseries histograms by binning the topic weighted posting times and measuring the distance to the uniform histogram. This distance is interpreted as a measure of each particular topics' temporal clustering, given the background (overall) rate of Tweeting. Ranking the results in descending order of distance, we show the range of distances in Fig. 1 and, in Figs. 2 and 3, a qualitative analysis of the 'furthest' four topics. We also include an analysis of the topic 'closest' to uniform for reference. Note that here we present only the results from December, though similar results have been generated for other months.

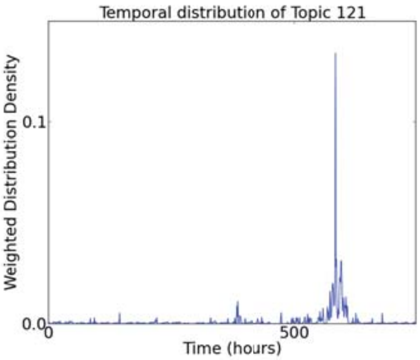
Figure 1 shows a small subset of topics on the left are considerably further from the uniform weighting than most other topics. Topics explored in depth (in Figs. 2 and 3) are the four leftmost points and the right most point on this plot. There is a clear change point in this plot on the left-hand side. While it is difficult to test the true cause leading to such as shape, we conjecture that the left-hand group of topics have some temporal linkage leading to more extreme EMD values and variation, while the majority of topics to the right of the change point do not have such a linkage.

4.4. *Application to twitter GPS data*

Keeping the timeseries histograms in mind, we would also like to know the topics with geographic histograms 'far' from the uniform histogram in space seen in Figure 2. Using the EMD, we can measure the distance from each topic's histogram to the uniform histogram seen in Figure 3. Ranking the results in descending order of distance, we show in Figs. 4 and 5 the results and, again, a short analysis of the four 'furthest' topic histograms, as well as a 'close' histogram for reference.

It is interesting to note that, in the geographic case, several topics are extremely far from the uniform distribution. As explored in the qualitative analysis, this may be attributed to user proclivity to Tweet certain things from only certain specific locations (e.g. local landmarks or the users' places of residence). The three furthest histograms (Topics 194, 80 and 166) have uni- or bi-modal distributions with very little spread. The fourth, however, is of particular interest due to its multi-modal nature and irregular shape. On the other end of the spectrum, we see that the closest to uniform topic includes words that could be used by all Twitter users. Several points on the far left of Fig. 2 the plot show extreme spatial localization. The topics explored in depth (in Figs. 4 and 5) are again the four leftmost points and the right most point on this plot.

Topic 121, Distance: 158.5699

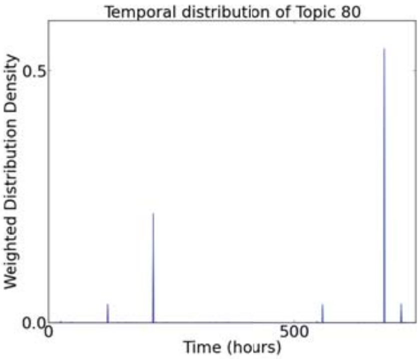


Top words:

1. merry
2. christmas
3. christmas[symbol]
4. mount
5. washington

Analysis: This topic encompasses Tweets about Christmas and posts about Mount Washington, which is both a local subdivision as well as a park with coinciding names. The location name is generated by Instagram.

Topic 80, Distance: 143.2101

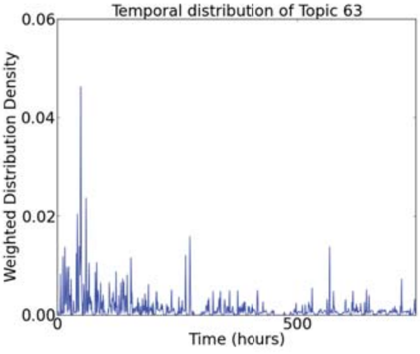


Top words:

1. rawr
2. ˆ0ˆ
3. kill
4. jurassic
5. dinosaur

Analysis: This topic is quite mysterious without user data but upon inspection appears to be a group of friends who use the word ‘rawr’, perhaps due to the Jurassic park movie. Their usage of the word is quite sparse.

Topic 63, Distance: 127.8254



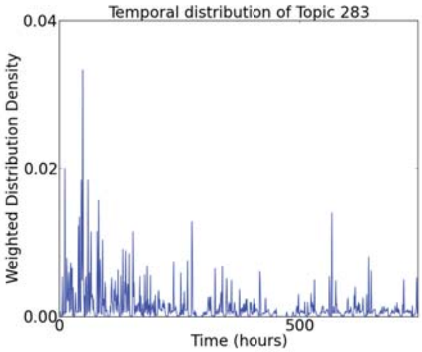
Top words:

1. 1183
2. unknown
3. injury
4. collision
5. traffic

Analysis: This topic encompasses posts by the California Highway Patrol, specifically for CHP code 1183 (Accident, no details). The pattern exhibited is consistent with weather patterns in Los Angeles, with the exception Christmas eve, which received heavy rain but low posting volume, implying a lower number of accidents.

FIG. 2. Display of three of the ‘furthest’ topic temporal histograms from the Uniform weighting using the EMD On the right of each section an interpretation of the topic is provided.

Topic 283, Distance: 118.9802

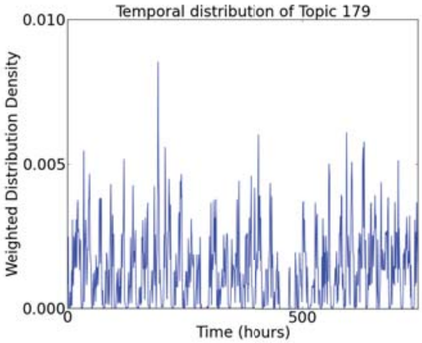


Top words:

1. 1182
2. injury
3. collision
4. traffic
5. vs

Analysis: This topic encompasses posts by the California Highway Patrol as well, specifically for incidents with CHP code 1182 (accident, property damage). It is parallel to the previous topic (63).

Topic 179, Distance: 2.6742

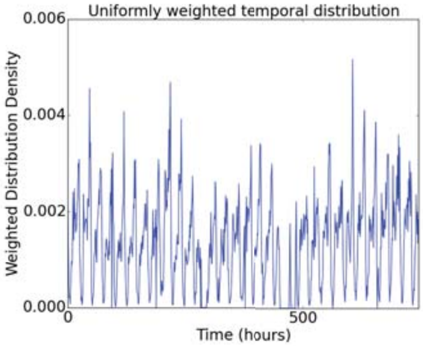


Top words:

1. got
2. present
3. [explicative]
4. card
5. nobody

Analysis: This topic is the closest to the uniform histogram. It somewhat describes the possible purchase of gifts and cards, with the mysterious inclusion of an explicative verb in past tense. This reflects the usage of 'got [explicative]'.

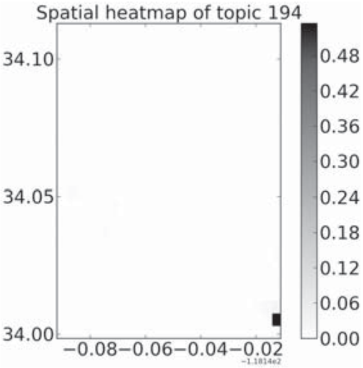
Uniformly Weighted Temporal Distribution of Tweets



We here display the uniformly weighted distribution of Tweets. There are clear cyclic patterns (on a 24-h scale, as well as possibly a weekly scale). Topic weight distributions that are relatively close to this distribution (as measured by the EMD) we interpret as being comparatively more uniform over time and thus less specific to particular events.

FIG. 3. Display of a 'far' distribution, the 'closest' distribution, as well as the uniform distribution (i.e. the zero-distance distribution), from top to bottom, respectively. On the right of each section an interpretation of the topic is provided.

Topic 194, Distance: 9.1704

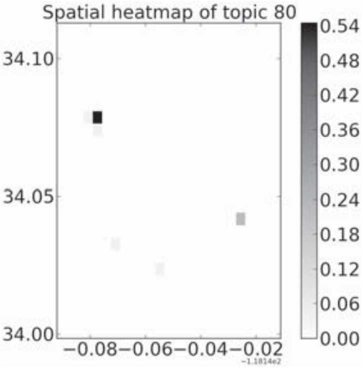


Top words:

1. citadel
2. outlets
3. commerce
4. shopping
5. others

Analysis: This topic appears to encompass Tweets from Citadel Outlet Malls, a shopping centre in Commerce, CA (a subdivision of Los Angeles).

Topic 80, Distance: 6.6391

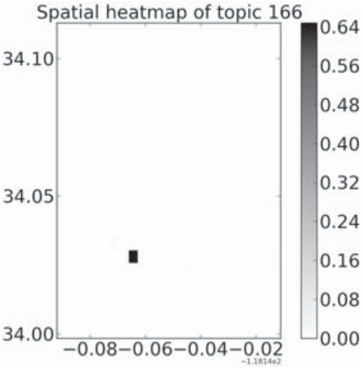


Top words:

1. rawr
2. ˆ0ˆ
3. kill
4. jurassic
5. dinosaur

Analysis: This topic is quite mysterious without user data but upon inspection appears to be a group of friends who use the word ‘rawr’, perhaps due to the Jurassic park movie.

Topic 166, Distance: 5.9912



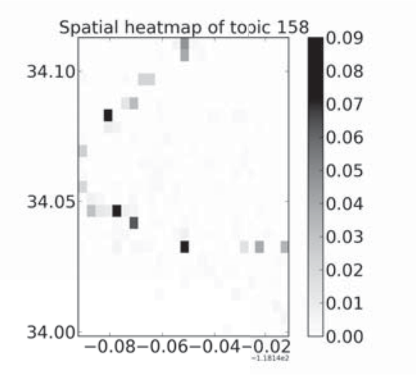
Top words:

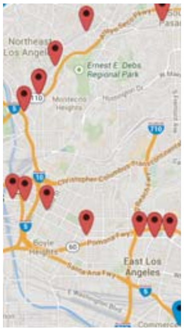
1. ty
2. gbu
3. jc
4. wanted
5. loving

Analysis: This topic also requires user data to interpret but upon inspection appears to be one man. He often uses the abbreviations ‘ty’, ‘gbu’ and ‘jc’. The active region appears to be his place of residence.

FIG. 4. We here display three of the ‘furthest’ topic spatial histograms from the Uniform weighting using the EMD. On the right of each section we provide an interpretation of the topic. The axes are longitude and latitude coordinates (the x-axis is relative to 118° W).

Topic 158, Distance: 3.7809



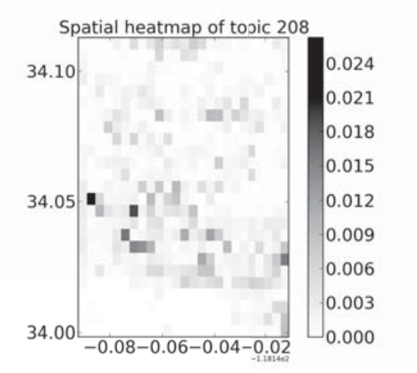


Top words:

1. tracking
2. graffiti
3. station
4. plaza
5. mariachi

Analysis:
This topic describes Tweets by a graffiti tracking service hired by the LA Metro Link. On the right hand side are the locations of the Metro Link stations in the area, which correspond with active regions. ‘Mariachi’ is one of the stations.

Topic 208, Distance: 0.2838



Top words:

1. check
2. dm
3. welcome
4. em
5. -_-

Analysis: This topic is the closest to the uniform histogram and is provided for reference. ‘dm’ is an abbreviation for Direct Message.

FIG. 5. Display of one ‘far’ topic spatial histograms and one ‘close’ topic spatial histogram, as measured by the EMD from the Uniform weighting. An interpretation of the topic is provided. The axes are longitude and latitude coordinates (the x-axis is relative to 118° W).

5. Point process models

In this section we construct the necessary definitions for our second method, providing brief discussion of their motivation and our specific usage. Results from this method are provided in Section 6.

5.1. *Hawkes process model*

A point process N is a random process where any realization consists of a collection of points typically representing the times and locations of events. The most basic of these processes is the stationary Poisson process in which events occur independently at a constant rate over an observed space-time window.

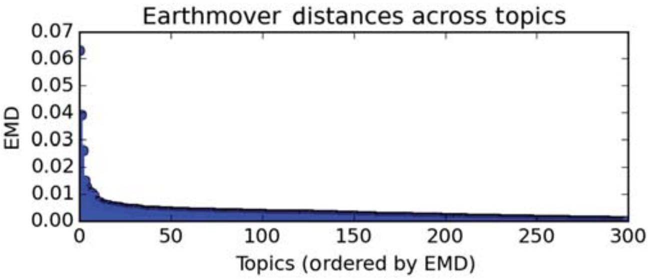


FIG. 6. EMDs for each spatial histogram.

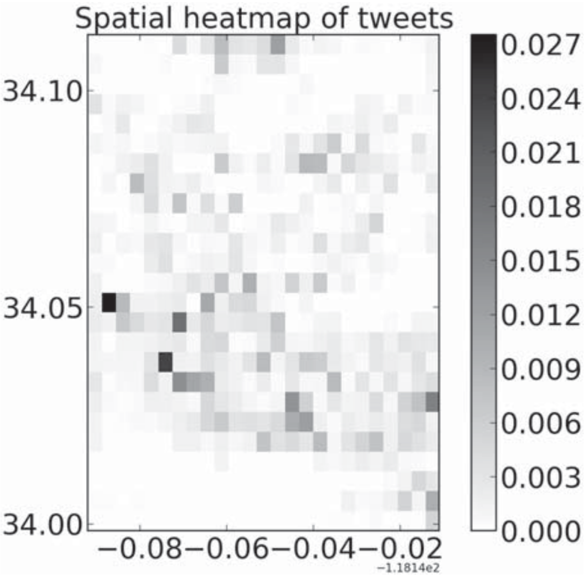


FIG. 7. The uniform histogram for geographic space.

Poisson point processes are characterized uniquely by their associated conditional rate λ, which is defined as the limited expected rate of the accumulation of points around a particular location and time (Daley & Vere-Jones, 2003).

In this work we focus on self-exciting point processes which describe sequences of events where the occurrence of one event increases the likelihood that another event occurs nearby in space and time. The Hawkes process (Hawkes, 1971) is one of the most important models of the conditional intensity for self-exciting point processes. This model was first applied to modelling earthquake occurrences through separate kernels for the background (mainshock) and triggering (aftershock) intensities. More recent applications include modelling spatial temporal crime rates (Mohler *et al.*, 2011), retaliatory acts of violence on a gang network (Stomakhin *et al.*, 2011) and e-mail traffic on a social network (Fox *et al.*, 2014).

For a sequence of Tweets of topic k, we model their associated time series $\{t_i^k : i = 1, \ldots, n_k\}$ as a Hawkes process with an exponential triggering function. The conditional intensity function $\lambda_k(t)$ for the rate at which Tweets occur in topic k is defined as:

$$\lambda_k(t) = \mu_k + \alpha_k \sum_{t_i^k < t} \omega_k \mathrm{e}^{\omega_k (t - t_i^k)}. \tag{5.1}$$

Here μ_k is the background rate for topic k, which can be interpreted as the occurrence rate of Tweets in topic k which are not triggered by other Tweets in topic k. The parameter α_k is the branching ratio of the process, which in context is the expected number of Tweets in topic k triggered by an arbitrary Tweet in topic k. The parameter ω_k governs the rate of decay, i.e. how quickly the overall rate λ_k returns to its background level μ_k after a Tweet occurs in topic k.

The exponential kernel was chosen because our topics come from a relatively short time interval. Sornette & Helmstetter (2003) suggest that for short time-scale topics the triggering kernel should obey an exponential decay function: $g(t) = \omega \mathrm{e}^{-\omega(t-t_k)}$.

In our analysis of the 1D Hawkes model, we mainly focus on the estimated branching ratio α_k since this parameter directly measures the amount self-excitation in the process and may be used to identify those topics where Tweets are highly clustered in time. We also compared the stationary temporal Poisson process with the exponential Hawkes process where the conditional intensity function of the former model is a constant.

5.2. *Marked spatio-temporal model*

The Hawkes process can be further extended to include both temporal and spatial information. Such a space-time process $N(t, x, y)$ is characterized via its conditional intensity $\lambda(t, x, y)$. For a sequence of N Tweets, we consider their sequence coordinates in space and time $(x_1, y_1, t_1), \ldots, (x_N, y_N, t_N)$ as such a process.

Point processes may also carry additional information beyond their location; these data are known as marks, and the corresponding processes are known as marked point processes. Here we carry the topic information as a mark, using notation similar to Mohler (2014), where the marks are used to denote different categories of crimes.

We consider the set of topics M believed to be precursory of one specific topic. For example, if we focus on the topic with descriptors ‘lakers game’, we consider topics that may be potential precursors (‘watch TV game’, ‘clippers lakers’). The topic label of a specific Tweet is indexed $z_{ij} \in \{0, 1\}$; $z_{ij} = 1$ if Tweet i is in topic j. The intensity of the topic specific process is now:

$$\lambda_k(t, x, y) = \mu(x, y) + \sum_{t_i < t} \sum_{j \in M} g(x - x_i, y - y_i, t - t_i, z_{ij}). \tag{5.2}$$

We use a triggering kernel which is specified as exponential in time and Gaussian in space:

$$g(x, y, t, z_{ij}) = z_{ij} \omega_k \theta_{j,k} \exp\left(-\omega_k t\right) \frac{1}{2\pi\sigma_k^2} \exp\left(-\frac{x^2 + y^2}{2\sigma_k^2}\right) \tag{5.3}$$

and a background rate estimated from all Tweets in the M topics:

$$\mu(x, y) = \sum_{t > t_i} \sum_{j \in M} z_{ij} \frac{\gamma_{jk}}{2\pi T \eta_k^2} \exp\left(-\frac{x^2 + y^2}{2\eta_k^2}\right). \tag{5.4}$$

In our model of the intensity function $\lambda_k(t, x, y)$ for topic k, $\theta_{j,k}$ is the expected number of Tweets in topic k triggered by an arbitrary Tweet in topic j; this is the main parameter characterizing the cross excitation rates between topics. Parameter σ_k is the standard deviation in distance among triggered Tweets, reflecting the spatial clustering of the topic. Parameter γ_{jk} gives the contribution of an event in a given topic j to topic k's background rate, ω_k is again the decay timescale and η_k is a background rate scaling parameter. T is the length of the observational window. The choice of these Gaussian functions in space allow for the derivation of the maximization step in the expectation-maximization (EM) algorithm for parameter estimation.

5.3. *Pre-processing and estimation*

In order to separate our Tweets by topic and to generate marks for our point processes, for topic encoding matrix W we normalize each row of the matrix. $W_{i,j}$ then represents the proportion of Tweet i consisting of topic j. We then threshold this matrix at a value of $\tau = 0.1$ and take any non-zero values as binary labels indicating membership in a topic. Note that some Tweets are effectively removed from our dataset as they have no assigned label. To estimate parameters, we use maximum-likelihood estimation via the EM algorithm of Veen & Schoenberg (2008).

5.4. *Extensions*

There are many natural extensions of our model that can be adapted to handle a variety of problems in future work. The topic model can be extended to a weighted semi-supervised model by applying Lee *et al.* (2010) for classification tasks, user specified topics or to weight rare topic classes. To deal with rare events, Vilalta & Ma (2002) and Weiss & Hirsh (2000) can be used as a predictive model across both rare and common time series patterns or to search for rare events.

6. Results and analysis

In this section, we present the results and analysis of the estimated Hawkes process models of the Twitter topics. We assess the goodness-of-fit of the models to the Twitter data with the Akaike information criterion (AIC) and non-parametric methods like the Kolmogorov–Smirnov (KS) test for the transformed times. We also interpret estimated parameters in the context of their respective topics.

6.1. *Temporal Hawkes model*

The AIC (Akaike, 1974) is defined as

$$\text{AIC} = 2\rho - 2l(\hat{\Omega}),$$

where ρ is the number of parameters of the model and $l(\hat{\Omega})$ is the maximum value of the log-likelihood function. AIC is a simple model selection criterion that encourages goodness-of-fit for a model (as given by likelihood) while penalizing the number of parameters, which serves as a measure of complexity. A smaller AIC value implies the model is a better fit.

As an initial validation of our model, we compute AIC scores for both a stationary Poisson model and a Hawkes model. Note, the intensity function for the stationary Poisson model is given by the constant rate $\lambda_k(t) = \mu_k$ for each topic k. Unlike AIC calculations for most models, AIC for point processes may be negative (Lewis *et al.*, 2012); the smaller (more negative) score denotes the preferred

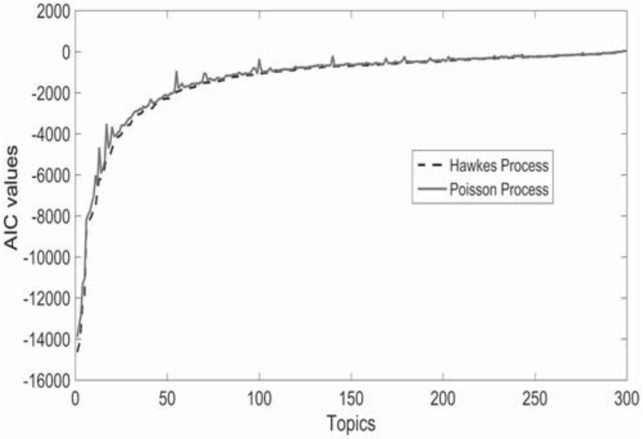


FIG. 8. AIC values for the Poisson and Hawkes models for each topic (labelled 1–300).

TABLE 2 *Number of parameters (ρ), maximum log-likelihood values ($l(\hat{\Omega})$) and AIC values ($2\rho - 2l(\hat{\Omega})$) for the temporal Hawkes and stationary Poisson models.*

	ρ	$l(\hat{\Omega})$	AIC
Stationary Poisson	300	194383.9	−388167.8
Temporal Hawkes	900	214483.2	−427166.3

model. Since the Hawkes model has more parameters than the Poisson model yet reduces to the latter in the case that any of the triggering parameters are zero, by calculating the AIC scores for each we can measure the amount to which a self-exciting model better fits the data. In every case for every topic the Hawkes model has a better AIC score (Fig. 8), though the margin varies by the amount to which a topic clusters. The AIC values are relative, but their overall magnitudes scale with the size of the data. While comparisons remain valid, the difference in the scores themselves will scale inversely with the size of the data. The maximum log-likelihood and AIC scores for the temporal Hawkes and stationary Poisson models summed over all 300 topics are presented in Table 2. This table shows that the Hawkes model performs significantly better than stationary Poisson over all Twitter topics according to the AIC.

Another goodness-of-fit diagnostic considered in Ogata (1988) is the transformed time $\{\tau_i^k\}$, which may be defined for each topic k as

$$\tau_i^k = \int_0^{t_i^k} \lambda_k(t)\,\mathrm{d}t. \tag{6.1}$$

If the conditional intensity is the true model used to generate the data then the transformed times follow a stationary Poisson process with rate 1 (Meyer, 1971). Hence, the inter-event times $\tau_i^k - \tau_{i-1}^k$ follow an exponential distribution, and consequently $U_i^k = 1 - exp\{-(\tau_i^k - \tau_{i-1}^k)\}$ follows a uniform distribution

TABLE 3 *Estimated parameters and KS test p values for the temporal Hawkes model for some select topics.*

No.	Top words	$\hat{\mu}$	$\hat{\alpha}$	$\hat{\omega}^{-1}$(day)	p value
1	‘ca’ ‘angeles’ ‘commerce’ ‘alhambra’ ‘monterey’ ‘jack’	18.97	1.60	0.053	3.8e-10
46	‘white’ ‘center’ ‘medical’ ‘memorial’ ‘lab’ ‘clinical’	8.25	0.13	0.00002	0.25
98	‘cold’ ‘af’ ‘outside’ ‘warm’ ‘weather’	7.88	0.60	0.059	0.88
251	‘game’ ‘clipper’ ‘laker’ ‘basketball’ ‘fan’ ‘video’	4.99	0.65	0.0567	0.83
294	‘@’ ‘photo’ ‘posted’ ‘hq’ ‘pic’ ‘bridge’	8.65	0.90	0.040	0.97
96	‘chico’ ‘fluffice’ ‘ice’ ‘rt’ ‘sexy’ ‘fan’	9.10	0.19	0.002	0.16
12	‘new’ ‘york’ ‘berrics’ ‘year’ ‘eve’ ‘twitcon’	9.25	0.81	0.0213	0.034
234	‘rawr’ ‘dinosaur’ ‘jurassic’ ‘seen’ ‘park’	0.55	0.36	4.15	0.78
—	All twitter data	6.29	0.99	0.0065	5.1e−21

over $[0, 1)$. Any deviation of $\{U_i^k\}$ from the uniform distribution corresponds to some feature in the data which is not well captured by the estimated model.

In Table 3 we present the p values from the Kolmogorov–Smirnov test comparing $\{U_i^k\}$ to the uniform distribution for some select topics k. A large p value (e.g.>0.05) indicates that the Hawkes model is well fit to the data, while a small p value (e.g. <0.05) indicates some feature of the data which is not well captured by the estimated model. Topics such as ‘cold af outside’ and ‘@ photo posted’ appear to fit well since the corresponding p values are greater than 0.05. Intuitively, we expect topics about the weather or posting photos to generate Tweets that are temporally clustered and thus fit well to the self-exciting model. In a few exemplar cases, the Hawkes model is less valid; e.g., the ‘ca angeles commerce’ and ‘new york berrics’ topics have small p values. For these topics we may be modelling noise or Tweets that generally do not cluster or possess self-exciting characteristics.

The last row of Table 3 shows the result of fitting the Hawkes model to the entire Twitter dataset (not conditioned on topics). The small p value indicates that a simple Hawkes model with three parameters cannot capture all the complexities in the entire dataset. Indeed, by classifying Tweets into their respective topics the Hawkes model is better fit and more adequately captures the temporal clustering in the data.

Lastly, Fig. 9 reveals that the p values for the Hawkes model are generally much larger than stationary Poisson for each topic. Moreover, 95.3% of the p values for the Hawkes models are greater than 0.05, indicating that this model is a good fit for most topics (in comparison, only 6.3% of the p values for the Poisson models are greater than 0.05).

Note that Figs. 8 and 9 show a much more substantial difference between the fitted Hawkes and Poisson models in terms of KS p values than AIC scores. However, this is perhaps not surprising since the AIC and KS test for the transformed times are two entirely different diagnostics: AIC is a

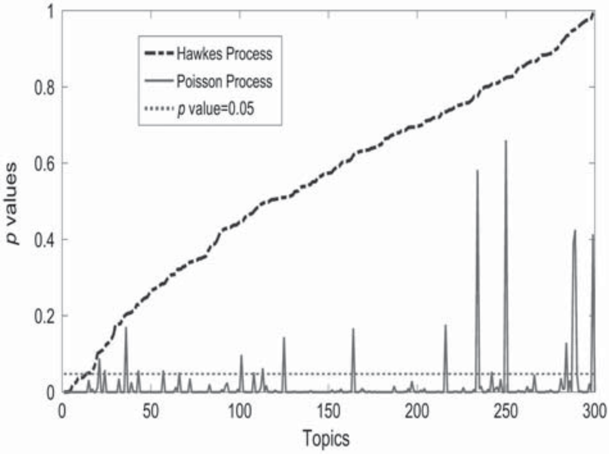


FIG. 9. KS test p values for the Poisson model and Hawkes model (dotted line) for each topic (labelled 1–300).

likelihood-based statistic used to compare nested models of varying complexity, while the KS p values are used to test whether the distribution of the transformed times deviates significantly from a uniform distribution. The two diagnostics also both indicate that the self-exciting model is a better fit to the data. Moreover, the results for the KS test suggest that there is a lot of clustering in the temporal point process data that is not being accounted for by the stationary Poisson model.

6.2. *Strongly branching topics*

Table 3 lists the estimated parameters of the Hawkes process models for some select topics. The $\hat{\alpha}$ branching ratio, equal to the estimated mean number of Tweets triggered per Tweet, is of particular interest. For instance, the '@ photo posted' topic is highly clustered in time since for every 100 tweets sent in this topic the estimated mean number of triggered Tweets is 90, and it takes on average 57 min for a Tweet in this topic to trigger another Tweet. Similarly, tweets in topics about the weather ('cold af outside') or sports ('game clipper laker') have strong estimated branching ratios and quick estimated average triggering times. By comparison, the topics 'white center medical' and 'rawr dinosaur jurrasic' have much weaker branching ratios.

6.3. *The response time of topics*

ω^{-1} represents the expected amount of time for a Tweet about one topic to trigger another Tweet of the same topic. Ranking ω^{-1} reveals topics with quick response times. For example, topic 96 is about fluff ice in East LA. This topic gives rise to immediate responses, since many individuals are familiar with this topic. Table 3 shows that there can be an order of magnitude difference in the decay rates for different topics. For instance, the expected response time for topic 234 about Jurassic Park is 4.15 days, while the expected response time for topic 251 about basketball is about 1.36 hours. This substantial difference in response times may correspond to the popularity of these topics, since a tweet about a current Lakers game is more likely generate quick responses than one about Jurassic Park.

6.4. *Marked spatio-temporal Hawkes model*

We again directly interpret the parameters of the Hawkes model fit to the data. As described in Section 5, σ shows the degree to which a topic clusters. We can, as in Section 4, directly rank these coefficients and investigate the extrema topics; e.g., the most spatially clustered topic is 'citadel outlets commerce' with $\sigma = 0.0006$ (which agrees with our results in Section 4) while the least spatial clustered topic with $\sigma = 0.0014$ is 'favorite seriously sad'.

Also described in the previous section is the parameter $\theta_{j,k}$, which, for each intensity function $\lambda_k(t,x,y)$, is the amount to which topic j triggers Tweets in topic k. Investigating $\theta_{k,k}$ is equivalent to investigating the self-excitation rate (this is similar to the parameter α in the 1D unmarked case). We again show only a few exemplar cases, as there are too many interactions to present (K^2 for K topics).

- M={Topic 123 ('end-of-world 2012'), Topic 113 ('happy sad')},

θ_{jk}	$j = (123)$	$j = (113)$
$k = 123$	0.13	0.00
$k = 113$	0.19	0.97

First, it is quite interesting to note the extremely high rate of self-excitation in the 'happy sad' topic. Second, discussion of the purported end of the world is a precursor to Tweets discussing 'happy sad'.

- M={Topic 127 ('traffic la'), Topic 82 ('food traffic')},

θ_{jk}	$j = (127)$	$j = (82)$
$k = 127$	0.78	0.48
$k = 82$	0.00	0.08

Los Angeles traffic is, unsurprisingly, a self-exciting topic, but the discussion of food and traffic is a strong precursor to a simple discussion of traffic. This may be due to the topic of food and traffic being semantically a subset of the topic of traffic as a whole.

- M={Topic 193 ('game clipper laker'), Topic 90 ('laker watching tv')},

θ_{jk}	$j = (90)$	$j = (193)$
$k = 90$	0.72	0.81
$k = 193$	0.00	1.95

First, note the extreme excitation rates of both topics; these are clearly well-clustered topics temporally. Discussion of the Lakers game informs on possible discussion of a Lakers–Clippers game.

Finally, we can investigate these interactions on a wider scale. We present a small example situation of four topics about the Lakers or related games, two topics about holidays and four topics about basketball in general. The resulting excitation coefficients are presented in Fig. 10, where darker means a stronger coefficient.

The results show that one type of holiday conversation is a strong precursor to discussion of basketball in almost every topic studied, but, appropriately, basketball does not provoke much conversation about the holidays.

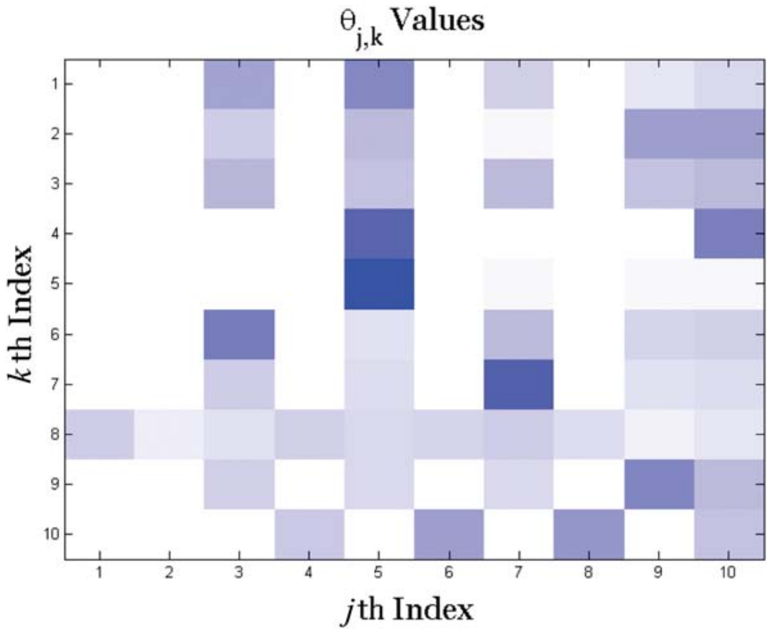


FIG. 10. θ_{jk} for topics 1 to 10, $M = 1, 2, \ldots, 10$. In order from left to right, the first four topics contain content about the Lakers, the next two contain holiday related content (in this case Christmas and New Years) and the next four contain content about basketball. Here, the column index denotes the 'preceeding' topic and the row index denotes the 'succeeding' topic. Darker cells indicate stronger coefficients.

7. Conclusions and discussion

In this article, we propose two methods for the analysis of generic topic models on corpora of text with spatio-temporal information. The first applies the EMD to topic histograms in order to discover topics that have abnormal structure in comparison with the background rate. The second measures clustering by self-excitation and then is extended to measure cross-excitation rates. We present results of both methods on a Twitter data set collected from East Los Angeles over a 10-month span, demonstrating their viability and usefulness. In particular, the first method immediately selects temporally and spatially clustered topics, where the clusters do not have a particular shape or distribution. The second method successfully recovers hidden interactions between topics which provides deeper insight into the underlying temporal and spatial structure of the data.

Funding

University of California, Los Angeles; Claremont McKenna College; W. M. Keck Foundation; University of California Lab Fees Research Grant [12-LR-236660]; National Science Foundation [DMS-1417674, DMS-1045536], Office of Naval Research [N00014-16-1-2119]; Air Force Office of Scientific Research-Multidisciplinary University Research Initiative [FA9550-10-1-0569].

Acknowledgements

The authors would like to thank Cristina Lopez, Xiyang Luo, Zhaoyi Meng, and Alexandre Robicquet for all their work alongside the authors during the 2014 Summer REU California Research Training Program in Computational and Applied Mathematics where this work began. Thanks to Yu-Han Chang, Rajiv Maheswaran and their group from USC's Information Science Institute for providing the Twitter data. We would also like to thank the 2013 LAPD RIPS group for their insightful discussions and suggestions.

References

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Autom. Control*, **19**, 716–723.

Applegate, D., Dasu, T., Krishnan, S. & Urbanek, S. (2011) Unsupervised clustering of multidimensional distributions using earth mover distance. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, pp. 636–644.

Blundell, C., Beck, J. & Heller, K. A. (2012) Modelling reciprocating relationships with Hawkes processes. *Advances in Neural Information Processing Systems* (F. Pereira, C. J. C. Burges, L. Bottou & K. Q. Weinberger eds), Curran Associates, Inc., pp. 2600–2608.

Cataldi, M., Di Caro, L. & Schifanella, C. (2010) Emerging topic detection on twitter based on temporal and social terms evaluation. *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, MDMKDD '10. New York, NY, USA: ACM, pp. 4:1–4:10.

Daley, D. & Vere-Jones, D. (2003) *An Introduction to the Theory of Point Processes, Volume 1: Elementary Theory and Methods*, 2nd edn. New York: Springer.

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

Ding, Z.-y., Jia, Y., Zhou, B., Han, Y., He, L. & Zhang, J.-f. (2013) Measuring the spreadability of users in microblogs. *J. Zhejiang Univ. Sci. C*, **14**, 701–710.

Fox, E. W., Short, M. B., Schoenberg, F. P., Coronges, K. D. & Bertozzi, A. L. (2014) Modeling e-mail networks and inferring leadership using self-exciting point processes. *J. Am. Stat. Assoc.*, accepted author version posted online Feb. 2016.

Godin, F., Slavkovikj, V., De Neve, W., Schrauwen, B. & Van de Walle, R. (2013) Using topic models for twitter hashtag recommendation. *Proceedings of the 22nd International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee. pp. 593–596.

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

Grindrod, P. (2014) *Mathematical Underpinnings of Analytics: Theory and Applications*. Oxford: OUP.

Hawkes, A. G. (1971) Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, **58**, 83–90.

He, X., Rekasina, T., Foulds, J., Getoor, L. & Liu, Y. (2015) HawkesTopic: a joint model for network inference and topic modeling from text-based cascades. *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pp. 871–880.

Hong, L., Ahmed, A., Gurumurthy, S., Smola, A. J. & Tsioutsiouliklis, K. (2012) Discovering geographical topics in the twitter stream. *Proceedings of the 21st International Conference on World Wide Web*. ACM, pp. 769–778.

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

Java, A., Song, X., Finin, T. & Tseng, B. (2007) Why we twitter: understanding microblogging usage and communities. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*. ACM, pp. 56–65.

Kim, H. & Park, H. (2008a) Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM J. Matrix Anal. Appl.* **30**, 713–730.

Kim, J. & Park, H. (2008b) Sparse nonnegative matrix factorization for clustering.

Kwak, H., Lee, C., Park, H. & Moon, S. (2010) What is twitter, a social network or a news media? *Proceedings of the 19th International Conference on World Wide Web*, WWW '10. New York, NY, USA: ACM, pp. 591–600.

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

Lee, H., Yoo, J. & Choi, S. (2010) Semi-supervised nonnegative matrix factorization. *Signal Process. Lett., IEEE*, **17**, 4–7.

Lewis, E. & Mohler, G. (2011) A nonparametric EM algorithm for multiscale Hawkes processes. preprint.

Lewis, E., Mohler, G., Brantingham, P. J. & Bertozzi, A. L. (2012) Self-exciting point process models of civilian deaths in Iraq. *Security J.*, **25**, 244–264.

Meyer, P. (1971) Démonstration simplifiée d'un théoréme de Knight. *Séminaire de Probabiliés V Université de Strasbourg*. Lecture Notes in Mathematics, vol. 191, Springer: Berlin: pp. 191–195. Heidelberg.

Mohler, G. (2014) Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *Int. J. Forecast.* **30**, 491–497.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P. & Tita, G. E. (2011) Self-exciting point process modeling of crime. *J. Am. Stat. Assoc.* **106**, 100–108.

Muskulus, M. & Verduyn-Lunel, S. (2011) Wasserstein distances in the analysis of time series and dynamical systems. *Physica D*, **240**, 45–58.

Noulas, A., Scellato, S., Mascolo, C. & Pontil, M. (2011) An empirical study of geographic user activity patterns in foursquare. *ICWSM*, **11**, 70–573.

Ogata, Y. (1988) Statistical models for earthquake occurrences and residual analysis for point processes. *J. Am. Stat. Assoc.*, **83**, 9–27.

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

Saha, A. & Sindhwani, V. (2012) Learning evolving and emerging topics in social media: a dynamic NMF approach with temporal regularization. *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*. ACM, pp. 693–702.

Salton, G. & McGill, M. J. (1983) Introduction to modern information retrieval.

Shirdhonkar, S. & Jacobs, D. W. (2008) Approximate earth mover's distance in linear time. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, vol. **0**, pp. 1–8.

Sornette, D. & Helmstetter, A. (2003) Endogenous versus exogenous shocks in systems with memory. *Physica A*, **318**, 577–591.

Stomakhin, A., Short, M. & Bertozzi, A. (2011) Reconstruction of missing data in social networks based on temporal patterns of interactions. *Inverse Problems*, **27**.

Veen, A. & Schoenberg, F. P. (2008) Estimation of space–time branching process models in seismology using an EM-type algorithm. *J. Am. Stat. Assoc.* **103**, 614–624.

Ver Steeg, G. & Galstyan, A. (2012) Information transfer in social media. *Proceedings of the 21st International Conference on World Wide Web*. ACM, pp. 509–518.

Vilalta, R. & Ma, S. (2002) Predicting rare events in temporal domains. *2002 IEEE International Conference on Data Mining, 2002. ICDM 2003. Proceedings.*, IEEE, pp. 474–481.

Weiss, G. M. & Hirsh, H. (2000) Learning to predict extremely rare events. *AAAI Workshop on Learning from Imbalanced Data Sets*, PP. 64–68.

Williams, S. A., Terras, M. M. & Warwick, C. (2013) What do people study when they study twitter? Classifying twitter related academic papers. *J. Doc.* **69**, 384–410.

Woodworth, J., Mohler, G., Bertozzi, A. & Brantingham, P. (2014) Nonlocal crime density estimation incorporating housing information. *Philos. Trans. R. Soc. A*, **372**(2028).

Yin, Z., Cao, L., Han, J., Zhai, C. & Huang, T. (2011) Geographical Topic Discovery and Comparison. *Proceedings of the 20th International Conference on World Wide Web*. ACM, pp. 247–256.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H. & Li, X. (2011) Comparing twitter and traditional media using topic models. *Advances in Information Retrieval*. Springer, pp. 338–349.

Zhou, K., Zha, H. & Song, L. (2013) Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*. JMLR, pp. 641–649.