Topic Time Series Analysis of Microblogs

Eric Lai [*] ellai@uci.edu
Daniel Moyer[§†] moyerd@usc.edu
Baichuan Yuan[‡] ybcmath@zju.edu.cn
Eric Fox[§] eric.fox@stat.ucla.edu
Blake Hunter[§¶] blake.hunter@claremontmckenna.edu
Andrea L. Bertozzi[§] bertozzi@math.ucla.edu
Jeffrey Brantingham[§] branting@ucla.edu

Abstract

Social media data tends to cluster in time and space around events, such as sports competitions and local news-worthy phenomena. However, transforming raw, free-form, real time text into meaningful information remains a challenging task. Confounding factors include the massive volume of posted data, lack of reliable event information, hidden temporal trends, and the vastly diverse nature of content. In the present work, we examine spatio-temporal topic distributions and self-exciting time series models as applied to social media microblog data. We apply topic modeling using non-negative matrix factorization with sparsity constraints to discover prevalent topics as well as latent thematic word associations within topics. We then present two methods for mining interesting spatio-temporal dynamics and relations among topics, one that compares the topic distributions directly, and another that models topics over time as temporal or spatio-temporal Hawkes process with exponential trigger functions. This second method allows identification of self-exciting topics and reveals unique temporal and spatial relationships among them.

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. When viewed as a time series, some of these topics might be observed in Tweets purely at random (topics associated with teenage romance perhaps), or on some periodic basis (topics about rush hour traffic, local weather, or a popular event). Still others, however, exhibit patterns quite different from baseline Twitter usage. Major holidays, one-time fads and social events, and pseudo-periodic events such as sport matches may be expected to produce anomalous distributions of Tweets with respect to the overall time series of Twitter.

Restricting Twitter to the geo-tagged Tweets, we might find a similar situation; while the entire corpus exhibits quite complex structures, some topics may be localized to certain areas and others may be distributed more globally. Tweets on a specific topic that cluster spatially, temporally or both might be of interest to analysts, marketers, researchers, law enforcement, and government agencies. The problem becomes one of identifying such interesting topics automatically from among the thousands to millions of topics observed in collections of microblog posts. The first half of our paper describes a method for finding these topics of interest.

Twitter topics may also have temporal or spatio-temporal relationships. 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 can inform on the incidence rate of Tweets from another [5, 28]. 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. This is indicative of a network structure of microblog topics, where edges represent the predictive power of one topic for another. Recovery of this latent network structure is discussed in the latter portion of this paper.

[*] University of California, Irvine
[†] University of Southern California
[‡] Zhejiang University
[§] University of California, Los Angeles
[¶] Claremont McKenna College

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
OCT 2014		**00-00-2014 to 00-00-2014**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Topic Time Series Analysis of Microblogs	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California, Los Angeles,Department of Mathematics,Los Angeles,CA,90095**	8. PERFORMING ORGANIZATION REPORT NUMBER **CAM14-76**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Social media data tends to cluster in time and space around events, such as sports competitions and local news-worthy phenomena. However, transforming raw, free-form, real time text into meaningful in- formation remains a challenging task. Confounding factors include the massive volume of posted data lack of reliable event information, hidden temporal trends, and the vastly diverse nature of content. In the present work, we examine spatio-temporal topic distributions and self-exciting time series models as applied to social media microblog data. We apply topic modeling using non-negative matrix factorization with sparsity constraints to discover prevalent topics as well as latent thematic word associations within topics. We then present two methods for mining interesting spatio-temporal dynamics and relations among topics one that compares the topic distributions directly, and another that models topics over time as temporal or spatio-temporal Hawkes process with exponential trig- ger functions. This second method allows identi cation of self-exciting topics and reveals unique temporal and spatial relationships among them.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**9**	

Here we consider 500,000 geolocalized Twitter messages from the Los Angeles area over a ten-month period. The Tweets are timestamped and geo-tagged (geographical location information from the user attached to the Tweet). We cleaned the Tweets by removing stop words and correcting misspellings, and then converted each one into an unordered histogram of words using a bag of words model [13], in which the order of words does not matter. Personal information including the user's handle was disregarded.

We present two methods for the automated discovery of such topics generated by Non-negative Matrix Factorization, one based on the Earth Mover's Distance and the other based on a self-exciting point process model.

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 behaviors also tend to cluster in time. The clustered, bursty nature of human behavior 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 behavior, forms dynamic spatio-temporal hotspots [7, 16, 18, 30].

2.2 Microblogs and Related Topic Models: Twitter as a source of data for academic study has been in use since approximately 2007 [10], 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 [29]). 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 [14]. Analysis of the text content includes both general models as well as Twitter specific models [9, 32].

Of particular note are two sub-classes of topic model, dynamic models [2], and geospatial models [8, 31]. For models in these two sub-classes, the set of topics is generated with respect to a prior distribution. The prior is dependent on the frequency of recent observations within topics, and/or the geo-spatial clustering of observations within topics. This prior biases the selection of topics, and the set of selected topics cannot be expected to describe the data fully (indeed, the authors do not claim such a thing), for there are topics which are equally expressed across time and space with respect to the overall volume of Tweets. For example, restricting our data to Los Angeles, it is reasonable to assume all users will experience traffic, and the frequency of Tweets should therefore be widely distributed across space and time (though not uniformly).

2.3 Spatial and Temporal Anomaly Detection: Directly related to our first method is a work by Applegate et al. [1]. The authors consider only usage data without content, applying an approximate Earthmover's Distance described in [26] 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 the approximate Earthmover's distance from regular histograms to any graphical structure (most importantly cyclic graphs), and then provides an analysis of the resulting clusters. Our first method is the application of a similar distance to spatio-temporal distributions of topics.

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. [32] use a Twitter generative text model based on LDA, then match Topics between the generated Twitter model and the New York Times. While Zhao et al. do not investigate linkages between topics in the same corpora, similar to the present work the authors investigate triggers between the two corpora, though not in a point process context. We extend this idea to the point process framework in section 5 and 6.

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 Non-Negative Matrix factorization 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 superficially here.

3.1 Pre-Processing: As found in [21], [6], and [9], 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. These we concatenate 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 [23]. 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 Non-negative Matrix Factorization (NMF) [15], 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 (KL) divergence instead of the Frobenius norm, NMF has been shown to be equivalent to Probabilistic Latent Semantic Indexing [4], 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 [22] [12]. This encourages sparsity and somewhat prevents overfitting. Our specific objective is given below:

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

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

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. Thus, by sorting the entries of the row we can form ranked lists of words describing the topic. Furthermore, 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. In this paper we use the active set method developed by Kim et al. [11].

4 Earthmover's Distance

In this section we first define the Earthmover's Distance (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.

4.1 Definition of the Earthmover's Distance: Let P and Q be discrete distributions:

$$P = \{(p_1, w_{p1}), \ldots, (p_N, w_{pN})\},$$
$$Q = \{(q_1, w_{q1}), \ldots, (q_M, w_{qM})\},$$
$$\sum_{i=1}^{N} w_{pi} = 1 \quad \text{and} \quad \sum_{i=1}^{M} w_{qi} = 1.$$

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:

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

We define the Earthmover's Distance (EMD) as

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

as seen in [19]. More intuitively, if P and Q were piles of dirt, the Earthmover's Distance 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 paper 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 1-dimensional 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):

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

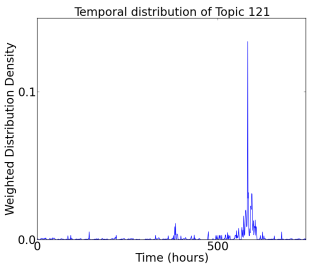
This is readily interpreted as the binned distribution of Tweets in L, reweighed by their topic encodings and normalized so that the bins sum to one. We also define the "uniform" weighting of the Tweets, which we refer to as the uniform histogram; note that this is **not** a Uniform distribution over 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 Earthmover's Distance originally formulated by Shirdhonkar and Jacobs [26] 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.

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. Ranking the results in descending order of distance, we show in Table 1 the results. Note that here we present only the results from December, though similar results have been generated for other months.

4.4 Application to Twitter GPS Data: Keeping the above histograms in mind, we would also like to know the topics with geographic histograms "far" from the uniform histogram in space. Using the EMD, we have a measure of this distance, so we can measure the distance from each topic's histogram to the uniform

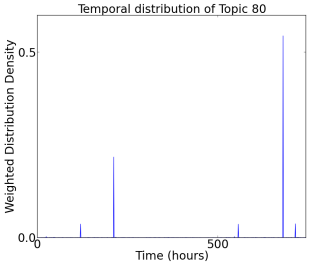
Topic 121, Distance: 158.5699



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.

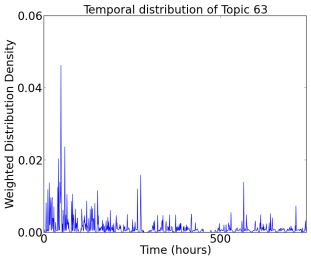
Topic 80, Distance: 143.2101



Top words:
1. rawr
2. ^0^
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.

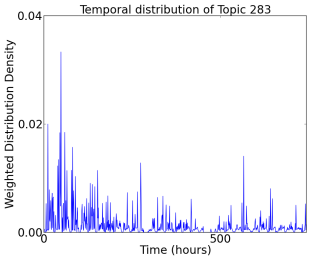
Topic 63, Distance: 127.8254



Top words:
1. 1183
2. unknown
3. injury
4. collision
5. traffic

Analysis: This topic encompasses automated posts by the California Highway Patrol, specifically for incidents with 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.

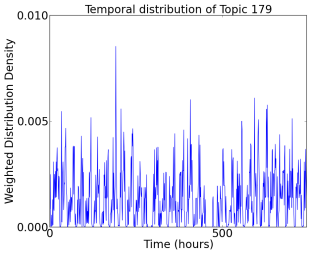
Topic 283, Distance: 118.9802



Top words:
1. 1182
2. injury
3. collision
4. traffic
5. vs

Analysis: This topic encompasses automated 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 one of 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]".

Table 1: A variety of topic histograms of Tweet density over time. The top four are the four furthest histograms from the uniform histogram, and the bottom is the closest.

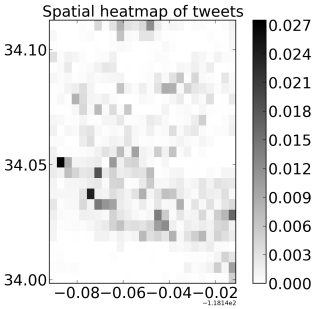


Figure 1: The uniform histogram for geographic space.

histogram. Ranking the results in descending order of distance, we show in Table 2 the results and a short analysis of the four furthest topic histograms, and a close histogram for reference. 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.

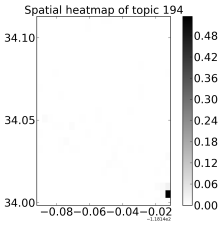
5 Point Process Models:

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 The Temporal Point 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 [3]. Each point process is characterized uniquely by its associated conditional rate λ, which is defined as the limited expected rate of the accumulation of points around a particular location. The most basic of these processes is the stationary Poisson Process, in which events are independent of each other, and occur at a uniform rate.

Non-stationary Poisson Processes are a generalization of the Poisson Process to non-constant rates. A further extension of this model is the removal of the independence assumption of the events; in particular, one important extension is the allowance for one event to either excite or dampen the probability of immediately observing another event. Here we use the Hawkes process, which takes some independent process as a background rate, and then makes the assumption that one event will trigger another with some decaying probability to be added on top of the background rate. Clearly this is dependent not only on the background rate (which need not be stationary) but also the choice of trigger function. The Hawkes process has been used

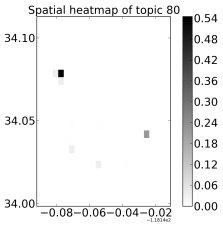
Topic 194, Distance: 9.1704



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 center in Commerce, CA (a subdivision of Los Angeles).

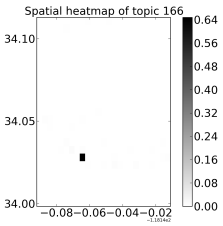
Topic 80, Distance: 6.6391



Top words:
1. rawr
2. ^0^
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.

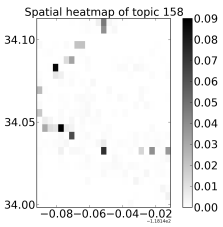
Topic 166, Distance: 5.9912

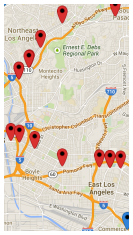


Top words:
1. ty
2. gbu
3. jc
4. wanted
5. loving

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.

Topic 158, Distance: 3.7809

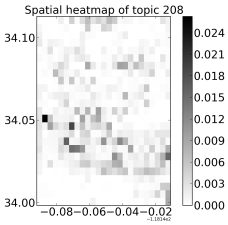




Top words:
1. tracking
2. graffiti
3. station
4. plaza
5. mariachi

Analysis: This topic describes Tweets by a graffiti tracking service hired by the LA Metro Link. On the righthand 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.

Table 2: A variety of topic histograms over space. The top four are the four furthest histograms from the uniform histogram, and the bottom is the closest. The axes are longitude and latitude coordinates (the x-axis is relative to 118° W).

in the past to model earthquakes [25] and financial markets, as well as crime [18].

For a sequence of Tweets of topic k, we consider their associated time series $t_1, \ldots, t_n$ as a realization of a Hawkes process $N_k(t)$ using an exponential trigger. The conditional intensity function $\lambda_k(t)$ of the one dimensional Hawkes process for topic k is defined as [17]:

$$\lambda_k(t) = \mu_k(t) + \alpha_k \sum_{t_n < t} \omega_k e^{-\omega_k(t-t_n)}.$$

Here, μ_k is the background rate for topic k (i.e. Tweets which are not triggered by other Tweets). α_k is the expected number of Tweets in topic k triggered by a Tweet in topic k, also known as the branching factor. ω_k is a parameter controlling 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. In our analysis of the one-dimensional Hawkes model, we mainly focus on the estimated α with respect to its topic and context.

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 $N(t, \vec{x})$ is characterized via its conditional intensity $\lambda(t, \vec{x})$. For a sequence of Tweets, we consider their sequence coordinates in space and time $(\vec{x}, t)_1, \cdots, (\vec{x}, t)_n$ 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 the similar notation to Mohler [18] 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 $z_{ij} \in \{0, 1\}$. The intensity of the topic specific process is now:

$$\lambda_k(t, \vec{x}) = \mu(\vec{x}) + \sum_{t > t_i} \sum_{j \in M} g(\vec{x}, \vec{x}_i, t, t_i, z_{ij}).$$

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

$$g(\vec{x}, \vec{x}_i, t, t_i, z_{ij}) = z_{ij}\omega\theta_{j,k} \exp\left(-\omega(t - t_i)\right) \times \frac{1}{2\pi\sigma^2} \exp\left(\frac{-\|\vec{x} - \vec{x}_i\|_2}{2\sigma^2}\right)$$

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

$$\mu(\vec{x}) = \sum_{t > t_i} \sum_{j \in M} z_{ij} \frac{\gamma_j}{2\pi T\eta^2} \exp\left(\frac{-\|\vec{x} - \vec{x}_i\|_2}{2\eta^2}\right).$$

The intensity function $\lambda_k(t, \vec{x})$ for topic k, $\theta_{j,k}$ is the number of Tweets in topic k triggered by an Tweet in topic j, and is the main parameter characterizing the relationship between the two topics. σ is the variance in distance among triggered Tweets, reflecting the spatial clustering of the topic. γ_j is the contribution of an event in a given topic to the background rate (zero if $z_i = 0$), ω is again the decay timescale, and η is a background rate scaling parameter. T is the length of the observational window.

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, and take any non-zero values as binary labels. Here we use a threshold $\tau = 0.1$, chosen after tuning. Note that some Tweets are effectively removed from our dataset as they have no assigned label. To estimate parameters, we use the Expectation-Maximization (EM) algorithm by Veen et al [27].

6 Results and Analysis

In this section we present results and analysis of the Hawkes process fit to our Twitter data set. We first include a short discussion of the Akaike Information Criterion (AIC) with respect to their preference in this dataset for self-exciting models, as well as the Kolmogorov-Smirnov (KS) Test of transformed data. For each model we then interpret selected parameters in the context of their respective topics.

6.1 Temporal Hawkes Model: AIC values can be used to compare the performance of different models on a fixed dataset [20]. As an initial validation of our model, we compute AIC scores for both a Poisson model and a Hawkes model. Though for most models AIC is strictly positive, here we use the AIC formulation for point processes given by Lewis et al. [17], in which negative values are expected. 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, though the margin varies by the amount to which a topic clusters.

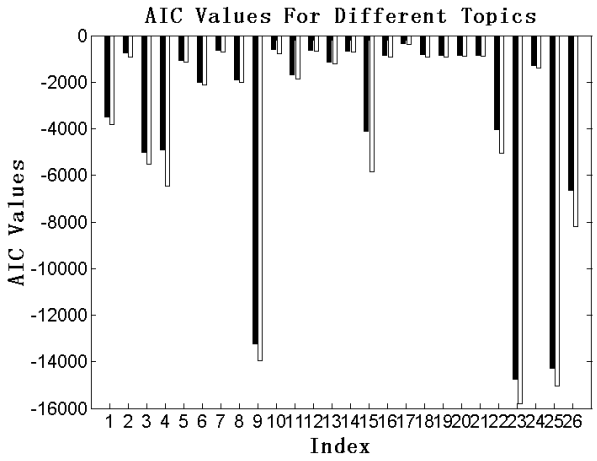


Figure 2: AIC values for Poisson and Hawkes models from topics 1 to 26. Negative is better. Note that this is only for a small subset of the topics.

We also calculate the p-values from the KS Test. In particular, after calculating the parameters of Hawkes process, we transform its estimated distribution to a process of constant unit rate. The KS test here measures the probability that the transformed time series distribution is drawn from the same distribution as a Poisson process. If the actual process is not a good fit, the transformation of the time series can be expected to not be Poisson distributed and thus deviate from the unit rate [24].

Topic	p-value	Top Words
88	0.8800	'cold' 'af' 'outside'
90	0.4297	'game' 'clipper' 'laker'
113	0.0060	'happy' 'sad' 'holiday' 'bday'
24	3.818e-10	'ca' 'angeles' 'commerce' 'alhambra'

Table 3: p-values of several topics.

The Hawkes model's validity decreases as the transformed rate deviates from the unit rate, so a higher p-value is better in this case. We only present a few exemplar cases, but clearly for some topics the Hawkes model is less valid, particularly when the time series is less clustered (in the case of topic 88, cold days could be considered a rare event in Los Angeles, and thus might be expected to cluster well).

Strongly Branching Topics: As well as interpreting general measures of "goodness-of-fit" and model validity we can directly interpret the estimated parameters. The α branching factor, equal to the expected number of Tweets triggered by an observed Tweet, is particularly interesting. In 4, we can see that topics simply describing a situation or action, or that are less coherent ("white center medical" and "chico fluff ice") hold much less predictive power than those used by Instagram to tag pictures, or the topics describing conversations between friends. The last topic is indeed this later case (See 2).

Table 4: Parameters of topics.

Topic Top Words	μ	α	ω^{-1}(day)
'white' 'center' 'medical'	8.25	0.13	0.00002
'@' 'photo' 'posted'	8.65	0.90	0.040
'cold' 'af' 'outside'	7.88	0.60	0.059
'chico' 'fluff' 'ice'	9.10	0.19	0.002
'rawr' 'dinosaur' 'jurassic' 'seen'	0.55	0.36	4.15

6.2 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; for example, the most spatially clustered topic is about "outlets shopping" with $\sigma = 0.0006$ (which agrees with our results in Section 4) while the least spatial clustered topic with $\sigma = 0.0014$ is about "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 one-dimensional 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 4 topics about the Lakers or related games, 2 topics about holidays, and 4 topics about basketball in general. The resulting excitation coefficients are presented below, where darker means a stronger coefficient.

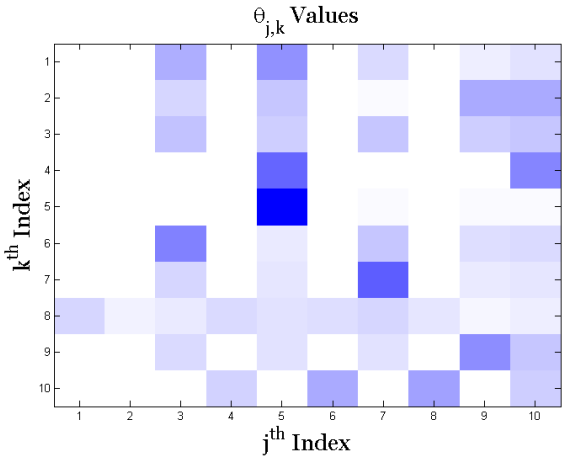


Figure 3: θ_{jk} for topics 1 to 10, $M = 1, 2, ..., 10$.

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.

7 Conclusions and Discussion:

In this paper, we propose two methods for the analysis of generic topic models on corpora of text with spatio-temporal information. The first applies the Earth-mover's Distance 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.

Acknowledgments

This work was supported in part by the University of California, Los Angeles; Claremont McKenna College; NSF Grant DMS-1045536 and AFOSR-MURI FA9550-10-1-0569. The authors would like to thank Cristina Lopez, Xiyang Luo, Zhaoyi Meng, 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 USCs 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

[1] David Applegate, Tamraparni Dasu, Shankar Krishnan, and Simon Urbanek. Unsupervised clustering of multidimensional distributions using earth mover distance. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 636–644. ACM, 2011.

[2] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, MDMKDD '10, pages 4:1–4:10, New York, NY, USA, 2010. ACM.

[3] Carlos Comas, Jorge Mateu, and Aila Särkkä. A third-order point process characteristic for multi-type point processes. *Statistica Neerlandica*, 64(1):19–44, 2010.

[4] Chris Ding, Tao Li, and Wei Peng. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Computational Statistics & Data Analysis*, 52(8):3913–3927, 2008.

[5] Zhao-yun Ding, Yan Jia, Bin Zhou, Yi Han, Li He, and Jian-feng Zhang. Measuring the spreadability of users in microblogs. *Journal of Zhejiang University SCIENCE C*, 14(9):701–710, 2013.

[6] Fréderic Godin, Viktor Slavkovikj, Wesley De Neve, Benjamin Schrauwen, and Rik Van de Walle. Using topic models for twitter hashtag recommendation. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 593–596. International World Wide Web Conferences Steering Committee, 2013.

[7] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[8] Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J Smola, and Kostas Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *Proceedings of the 21st international conference on World Wide Web*, pages 769–778. ACM, 2012.

[9] Liangjie Hong and Brian D Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88. ACM, 2010.

[10] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[11] Hyunsoo Kim and Haesun Park. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM Journal on Matrix Analysis and Applications*, 30(2):713–730, 2008.

[12] Jingu Kim and Haesun Park. Sparse nonnegative matrix factorization for clustering. 2008.

[13] Raymond Kosala and Hendrik Blockeel. Web mining research: A survey. *ACM Sigkdd Explorations Newsletter*, 2(1):1–15, 2000.

[14] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

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

[16] Erik Lewis and George Mohler. A nonparametric EM algorithm for multiscale Hawkes processes. *preprint*, 2011.

[17] Erik Lewis, George Mohler, P Jeffrey Brantingham, and Andrea L Bertozzi. Self-exciting point process models of civilian deaths in Iraq. *Security Journal*, 25(3):244–264, 2012.

[18] George Mohler. Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *International Journal of Forecasting*, 30(3):491–497, 2014.

[19] Michael Muskulus and Sjoerd Verduyn-Lunel. Wasserstein distances in the analysis of time series and dynamical systems. *Physica D: Nonlinear Phenomena*, 240(1):45–58, 2011.

[20] Yosihiko Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27, 1988.

[21] Daniel Ramage, Susan T Dumais, and Daniel J Liebling. Characterizing microblogs with topic models. *ICWSM*, 10:1–1, 2010.

[22] Ankan Saha and Vikas Sindhwani. Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 693–702. ACM, 2012.

[23] Gerard Salton and Michael J McGill. Introduction to modern information retrieval. 1983.

[24] Frederic Paik Schoenberg. On rescaled Poisson processes and the Brownian bridge. *Annals of the Institute of Statistical Mathematics*, 54(2):445–457, 2002.

[25] Frederic Paik Schoenberg, David R Brillinger, and Peter Guttorp. Point processes, spatial-temporal. *Encyclopedia of environmetrics*, 2002.

[26] Sameer Shirdhonkar and David W. Jacobs. Approximate earth mover's distance in linear time. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 0:1–8, 2008.

[27] Alejandro Veen and Frederic P Schoenberg. Estimation of space–time branching process models in seismology using an EM–type algorithm. *Journal of the American Statistical Association*, 103(482):614–624, 2008.

[28] Greg Ver Steeg and Aram Galstyan. Information transfer in social media. In *Proceedings of the 21st international conference on World Wide Web*, pages 509–518. ACM, 2012.

[29] Shirley A Williams, Melissa M Terras, and Claire Warwick. What do people study when they study twitter? classifying twitter related academic papers. *Journal of Documentation*, 69(3):384–410, 2013.

[30] JT Woodworth, GO Mohler, AL Bertozzi, and PJ Brantingham. Nonlocal crime density estimation incorporating housing information. *Philosophical Transactions of the Royal Society A*, 372(2028), 2014.

[31] Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang. Geographical topic discovery and comparison. In *Proceedings of the 20th international conference on World wide web*, pages 247–256. ACM, 2011.

[32] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338–349. Springer, 2011.