Talking Places: Modelling and Analysing Linguistic Content in Foursquare

Sandro Bauer, Anastasios Noulas, Diarmuid Ó Séaghdha, Stephen Clark, Cecilia Mascolo
Computer Laboratory, University of Cambridge, UK
firstname.lastname@cl.cam.ac.uk

***Abstract*—**
The advent of online social media and the growing popularity of sensor-equipped mobile devices have created a vast landscape of location-aware applications and services. This goldmine of data, including temporal and spatial information of unprecedented granularity, can help researchers gain insights into the behavioural patterns of people at a global scale. Here we analyse the textual content of millions of comments published alongside Foursquare user check-ins. For this, we extend a standard topic modelling approach so that it explicitly takes into account geographic and temporal side information. The framework is applied to Foursquare data and used to detect the dominant topics in the neighbourhoods of a city. In particular, we present the most prominent topics discussed by Foursquare users in New York, London, Chicago and San Francisco. We characterize the topics' spatial coverage and temporal evolution, and we also highlight some cultural idiosyncrasies. Finally, we evaluate the novel spatio-temporal topic model quantitatively. We believe that our model may be a useful tool for social scientists and application developers.

I. Introduction

Online Social Networks, such as Facebook or Twitter, and services relying on them have not only transformed people's communication habits, but they are also producing an unprecedented amount of data capturing the behavioural patterns of millions of users globally. While data generated automatically by digital means has its own inherent limitations, its sheer scale and granularity promise to open up research avenues unimaginable in earlier decades, when extensive surveys were the only way to obtain representative data.

Since the early days of such networks, users have been encouraged to establish friendship or follower relationships with their acquaintances, and to share multimedia content by broadcasting micro-blogging messages or posting on each other's wall. There was little information available, however, on the real-world *context* in which these messages were published, beyond a timestamp. For example, an enthusiastic comment on the music played in a user's surroundings might be much more relevant to other users if they know *where exactly* in town the user is writing from. This side information would have had to be provided manually in the past (and often was not), since accessing the web was only possible through desktop computers connected to landlines.

More recently, however, smartphones and tablet PCs have allowed users to access on-line social services from virtually everywhere. Many of these devices have their own GPS sensor, and so it is straightforward to collect explicit geographic side information, thereby allowing users to map their on-line content to places of interest directly. This idea was received with enthusiasm by many users, and hence the proportion of geo-tagged tweets is constantly increasing. The geo-social hype in recent years even led to the creation of separate mobile social networks, known as *Location-based Social Networks* (*LBSNs*). The most popular of these platforms, Foursquare, has attracted more than twenty million users so far [18], bringing the problem of *place recommendation* to the attention of both users and researchers.

There are large and growing literatures on mining mobility data associated with LBSNs on the one hand and textual data from social media on the other, but there has been relatively little research that takes into account the availability of both modalities (see Section V for an overview). Furthermore, applications like Foursquare associate a further metadata layer with real-world locations by providing category information, such as whether a location is a bar or a cinema.

We have collected a large dataset of Foursquare check-ins and accompanying textual content in order to evaluate the potential of natural language processing (NLP) techniques for exploring the content published in Location-based Social Networks. Specifically, the paper makes the following contributions:

- We describe a spatio-temporal extension to the well-known Latent Dirichlet Allocation (LDA) topic model [2]. While previous efforts have adapted LDA to either the temporal or the spatial domain [9], [26], we integrate both dimensions simultaneously. We apply our model in the context of Mobile Social Networks, an environment which presents a natural fit to our approach, since the spatio-temporal dynamics of user-generated comments can be modelled in order to sense the heartbeat of urban neighbourhoods.
- We investigate the ability of topic models to recover the topics discussed in a city by Foursquare users. Moreover, we rank these topics according to an intensity measure, observing that the most popular topics are similar across cities and correspond to well-known urban activities (work, food, nightlife etc.). We also demonstrate the existence of topics which provide insights into the cultural idiosyncrasies of cities in terms of human activity, the nature of their neighbourhoods and geographic characteristics. By relating the categories of Foursquare places to the inferred topics, we show how the situational context of

users affects the user-generated linguistic content.
- Further, in response to our analytical observation that the geographic spread of topics in a city may vary significantly, we characterize them using an entropy-based measure to detect such heterogeneities. We show that some topics in Foursquare are highly focused geographically, while others span larger geographic areas. In addition to looking at spatial variations, we analyse the evolution of topics over time, capturing the rhythms of user discussions across different hours and days of the week.

The analysis was conducted on a dataset comprising millions of Foursquare user check-ins collected over a period of six months in 2010, along with textual content published on Twitter. We focus on topical patterns observed in four large cities around the world (London, New York, San Francisco and Chicago). We think that our work may be a a useful tool for social scientists, as well as a resource for mobile applications such as place recommendation systems. Further possible application scenarios include content targeting for mobile users as well as urban planning, where it may seem appealing to exploit the expressive power of natural language to gain insights into the activities carried out in a city's various neighbourhoods.

The paper is organised as follows. We begin by describing both the Foursquare dataset we crawled and our spatio-temporal topic model in detail. Next, we provide a comprehensive evaluation of the model on Foursquare data and present the topics that have emerged in the four cities. Finally, we describe related work and conclude.

II. Foursquare Dataset

Foursquare[1] is a Location-based Social Network launched in 2009. Using a mobile web application, users are able to connect to their friends and share their whereabouts by *checking in* at virtual places which are mapped to real-world venues via their geographic coordinates. The service was originally designed as a game in which a user would become the *mayor* of a place if they were the person with the highest number of *check-ins* there. In the past few years, Foursquare has grown to become the large-scale social network we know today, with more than 20 million registered users. With such a massive user base, the network is now in a position to serve as a *place recommendation engine* which users rely on to explore neighbourhoods and discover new places.

When checking in to a Foursquare venue, a user may optionally publish a comment onto his Twitter timeline. We have collected a public dataset of such Foursquare-sourced tweets which covers a period of more than five months (from May 27th to November 2nd, 2010). In total we have collected approximately 35 million check-ins, which accounts for about 25% [17] of all Foursquare check-ins in this period. We have also acquired a dataset of $4,960,496$ venues from the Foursquare website. For each venue, there is a variety of properties available, such as its exact geographic coordinates, its category (*Library, Train Station* etc.), and locality information describing the city it belongs to.

[1] http://www.foursquare.com

City	4sq tweets	Tokens	Tokens*
London	54,179	420,972	166,668
New York	296,881	2,136,039	850,501
Chicago	104,237	753,893	300,739
San Francisco	74,642	536,273	216,458

TABLE I: Datasets produced for the four cities. The original input is Twitter messages sourced from Foursquare (**4sq Tweets**). We also provide the total number of tokens in the datasets before (**Tokens**) and after (**Tokens***) filtering out Foursquare-specific words (mayor, 4sq etc.).

In this work, we focus on four large cities with a high number of check-ins on Foursquare: New York, Chicago, San Francisco and London. Our goal is to analyse the linguistic content of Foursquare user comments in these cities by inferring a set of topics representing this content. As a pre-processing step, we shift the publication times of all tweets (given in UTC) to local time. From the tokens we use for training the topic models, we remove stopwords, very short and overly long tokens, sentence markers and emoticons, as well as a hand-crafted collection of Foursquare-specific sentence patterns (words such as *mayor, check-in* etc.). Detailed statistics on the datasets produced for the four cities can be found in Table I.

III. A Spatio-Temporal Topic Model

Latent Dirichlet Allocation (LDA) [2] is an unsupervised machine learning method for discovering thematic structure in text, based on patterns of co-occurrence between words. The basic version of LDA assumes that words of the text corpus being modelled are partitioned into a set of *documents*. These can be real-world documents, but extensions to the model have used different notions of document as well; for example, researchers working with Twitter data often aggregate all tweets written by a single user into a "superdocument" due to the brevity of individual messages. Each document d is associated with a probability distribution θ_d over K latent variable indices or *topics*, and each topic index z is associated with a distribution Ψ_z over words. Parameter estimation involves leveraging co-occurrence information within and across documents to assign a single topic index to each token in the corpus; the "meaning" of each topic emerges from the clustering behavior observed across tokens assigned to the same topic.

The Spatio-Temporal Topic Model presented here is different in that it uses multiple partitions of the data. Each token in the corpus is associated with a timestamp and a location; these have a similar status to documents in LDA as we associate topic distributions θ_t and θ_l with each temporal and spatial "chunk" of the corpus. The generative model we assume states that the choice of each word in the corpus depends on either the temporal properties of the message that contains it, the message's spatial properties, or neither. We hypothesise that some of the content words cannot be explained by the location or time of publication, but will be present across the whole corpus. We want the model to push these words into a dedicated background topic so that the remaining topics will reflect better the geographic and temporal patterns in the corpus. The

learning algorithm must decide which of these three options to associate with each token. Inspired by Chemudugunta et al. [3], we integrate a second layer of latent *switching variables* s whose values are set in parallel to the topic variables. This procedure allows for the interpretation of an observed message as the outcome of multiple real-world processes, each of which is assumed to be responsible for the appearance of *some* of the words in the tweet. In fact, we do not use tweet-specific topic distributions at all here due to the inherent shortness of tweets (up to 140 characters only). The generative story is summarised in Figure 1. A visualisation is given as a plate diagram (cf. Figure 2).

The temporal and spatial side information for each token is extracted in the form of timestamps and latitude/longitude coordinates. As these are continuous rather than discrete, we specify geographic and temporal "boxes" of tweets that are grouped together for the purpose of analysis. We do so by sub-dividing a given area (e.g. a large city) into a grid of sub-areas, where the number of areas is chosen according to the problem of interest and the amount of data available. The same applies to the temporal groupings: In some cases, it might be interesting to observe the variation of a number of topics throughout the day (regardless of weekly or monthly variations), whereas for some problems weekly changes of a topic's intensity could be more informative. By using our command-line tool which makes tuning these parameters easy, practitioners will be able to quickly inspect their data in various ways to extract meaningful patterns.

In order to infer the latent random variables z and s, we use standard Gibbs sampling techniques. Similar to Rosen-Zvi et. al. [13], we sample the two latent variables z_i and s_i jointly. The sampler must be modified to respect the fact that the choice of one of the two hidden variables (s_i) puts a hard constraint on the choice of the topic, i.e. $p(z_i = z_b|s_i = b) = 1$, $p(z_i \neq z_b|s_i = b) = 0$, $p(z_i = z_b|s_i \neq b) = 0$ and $p(z_i \neq z_b|s_i \neq b) = 1$, where z_b denotes the designated background topic.

The conditional distribution of a pair of latent variables z_i and s_i, given hyperparameters Ξ, α, β and the latent variable assignments z^{-i}, s^{-i} for all other tokens in the corpus, is updated as specified below (the conditioning variables are omitted for simplicity), where N is the total number of words and $c(x)$ is a counting function.

$$p(z_i, s_i|\ldots) \propto \begin{cases} \frac{c(l)+\Pi_s}{N+\sum_s \Pi_s} \cdot \frac{c(z=z_i, s=l, l=l_i)+\alpha_z}{c(s=l, l=l_i)+\sum_z \alpha_z} \cdot \frac{c_{wz_i}+\beta_{wz_i}}{c_{z_i}+\sum_w \beta_{wz_i}} & s_i = l \\ \frac{c(t)+\Pi_s}{N+\sum_s \Pi_s} \cdot \frac{c(z=z_i, s=t, t=t_i)+\alpha_z}{c(s=t, t=t_i)+\sum_z \alpha_z} \cdot \frac{c_{wz_i}+\beta_{wz_i}}{c_{z_i}+\sum_w \beta_{wz_i}} & s_i = t \\ \frac{c(b)+\Pi_s}{N+\sum_s \Pi_s} \cdot \frac{c_{wz_b}+\beta_{wz_b}}{c_{z_b}+\sum_w \beta_{wz_b}} & s_i = b, z_i = z_b \\ 0 & s_i = b, z_i \neq z_b \\ 0 & s_i \neq b, z_i = z_b \end{cases} \quad (1)$$

Draw a global switching variable prior distribution $\Pi \sim \text{Dir}(\Xi)$
Draw a distribution over words $\Psi_b \sim \text{Dir}(\beta)$ for the background topic
For each topic $k \in \{1, ..., K\}$:
    Draw a topic-specific distribution over words $\Psi_k \sim \text{Dir}(\beta)$
For each area $l \in \{1, ..., L\}$:
    Draw an area-specific distribution over topics $\theta_l \sim \text{Dir}(\alpha)$
For each time frame $t \in \{1, ..., T\}$:
    Draw a time-specific distribution over topics $\theta_t \sim \text{Dir}(\alpha)$
For each word occurrence $(i, l_i, t_i) \in \{1, ..., N\}$ where l_i, t_i are given geo-temporal side information:
    Draw a word-specific switching variable $s_i \in \text{Mult}(\Pi, 1)$, with $\Xi \in \{l, t, b\}$, where l, t and b stand for the geographic, temporal and background source respectively
Depending on the choice of s_i do:
If $s_i = b$:
    Draw a word $w_i \sim \text{Mult}(\Psi_b, 1)$
If $s_i = l$:
    Draw a topic $z_i \sim \text{Mult}(\theta_{l_i}, 1)$
    Draw a word $w_i \sim \text{Mult}(\Psi_{z_i}, 1)$
If $s_i = t$:
    Draw a topic $z_i \sim \text{Mult}(\theta_{t_i}, 1)$
    Draw a word $w_i \sim \text{Mult}(\Psi_{z_i}, 1)$

Fig. 1. Generative story of the spatio-temporal topic model

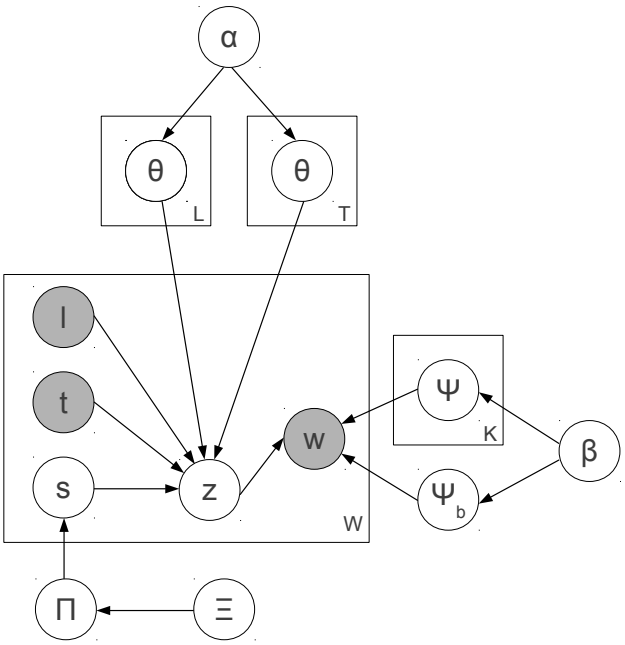


Fig. 2. Plate diagram of the Spatio-Temporal Topic Model. Latent variables are represented by white circles, whereas grey circles refer to observed variables. The designated background topic's distribution over words is labeled Ψ_b.

IV. Evaluation

We evaluate our spatio-temporal topic model on the textual data Foursquare users have published from venues in four large cities: London, Chicago, New York and San Francisco. The posts were published to Twitter from users' mobile phones alongside the check-ins. This way, they can be directly linked to the geo-coordinates of the corresponding Foursquare venue. We will first describe the parameterisation of our experiments and then go on to illustrate our main observations using a set of representative topics detected by the model, as well as their geographic and temporal characteristics.

london*	time	good	great	nice	waiting	place	back	love	free
work	back	week	coffee	working	today	office	busy	flat	latte
dinner	drinks	night	food	birthday	evening	tonight	party	watching	drink
food	amazing	love	chicken	happy	street	team	favourite	birthday	lovely
pint	beer	lovely	pizza	chips	drink	england	fish	local	wine
meeting	event	office	media	beautiful	session	business	social	building	today
train	home	back	heading	london	waiting	bound	station	trains	homeward
lunch	coffee	afternoon	today	sunday	brunch	lunchtime	meetings	food	salad
morning	breakfast	coffee	today	early	start	friday	meetings	monday	bacon
shopping	shop	buying	place	checking	bought	store	stop	shoes	champagne

TABLE II: Top 10 topics (in rows) inferred in London. The background topic is marked with a star.

time*	love	back	great	good	finally	free	waiting	stop	place
party	happy	birthday	bday	friends	place	drink	show	night	work
place	pizza	yummy	late	happy	favorite	cheese	delicious	chocolate	meal
work	coffee	back	today	good	iced	office	working	latte	busy
dinner	night	drinks	tonight	event	food	good	hour	home	sushi
working	meeting	today	weekend	friday	class	break	money	making	social
food	quick	picking	hungry	chicken	sushi	stuff	gettin	fresh	eating
event	meeting	week	york	great	hotel	team	media	room	checking
lunch	brunch	today	salad	sunday	sandwich	time	place	breakfast	burger
wine	drinks	amazing	restaurant	cocktails	friend	date	hanging	eating	dessert

TABLE III: Top 10 topics (in rows) inferred in New York. The background topic is marked with a star.

A. Experimental Setup

In each of the four cities, we choose an area of 256 km^2 and sub-divide it into 1,024 equal squares. Thus, each square unit is 500 by 500 metres long (typical walking distance). For conducting temporal analysis, we use five time slots representing *night* (0-5 h), *morning* (5-11 h), *noon* (11-14 h), *afternoon* (14-18 h) and *evening* (18-0 h).

For both our new model and ordinary LDA, we run the training procedure for 1,000 iterations. After a burn-in period of 200 iterations we re-estimate the α and β hyperparameters every 10 iterations using an iterative method proposed by [21]. We set the number of topics K to a default value of 100; using more topics than required to reveal the topical clusters present in the dataset does not affect the overall performance, as the bulk of the token mass will be moved into a limited number of topics only [20].

B. General Analysis of Topics

First of all, we examine what topics are most popular among Foursquare users, and to what subject matters they belong. The overall frequency of a topic in the corpus can be defined as the proportion of all tokens assigned to this topic by the model:

$$\text{intensity}_z = \frac{c(w|z)}{N} \tag{2}$$

A topic is a probability distribution over all the word types in the corpus. Tables II to V contain this data for the four cities mentioned above. We list the 10 most intensive topics inferred for each city (one per row), sorted according to the intensity values obtained from equation 2. Each topic is represented by its 10 most prominent words and the most prominent token is printed in bold.

Also, it is interesting to know whether these observations are coherent across cities and time spans, or if we can observe any characteristic differences. For this, we need to measure the topic intensity in a single square and a particular time frame. Analogously, we define the intensity of a topic z in area l and time frame t as the proportion of all tokens w in this area and time frame which have been assigned topic index z by the model:

$$\text{intensity}_{t,l,z} = \frac{c(w|t,l,z)}{c(w|t,l)} \tag{3}$$

By inspecting the tables we first note that the dedicated background topic produced by our model (to be found in the first row of the tables, respectively) is similar in all four cities, with the city's name itself being prominent in all background topics except in that for New York. The intensity of the background topic is very high (around 0.2) for all cities. This confirms our intuition that a substantial share of user content in LBSNs is unrelated to a user's current location or the time of publication.

Furthermore, we observe that the most popular topics (apart from the background topic) in all cities feature similar sets of words. Indeed, topics such as *food*, *party*, *dinner*, *morning*, *lunch* or *work/office* are not only observed across all cities, but they are also the most prominent ones. In addition to the noticeable semantic relationship between the tokens observed within a topic, we also observe the clear presence of the temporal dimension. In all cities, there are *night*, *morning* and *lunch* topics featuring words typical of a particular time of day. The recovery of these patterns is supported by explicitly taking into account the temporal dimension in the model, hence allowing us to capture not only what is being discussed somewhere, but at which times of the day/week as well. Note

that in the above examples, the temporal patterns observed are primarily time-of-day-specific because the time boxes we use aggregate time spans of several hours, regardless of the day of the week or any other temporal granularity. We have also experimented with day-of-the-week and monthly temporal groupings and similarly obtained characteristic patterns (not shown here), such as a weekend topic mentioning typical leisure activities as well as religious practice.

Other than linguistic patterns common to all the cities we used for our experiments, we have also observed topics unique to a specific city or district, especially in the case of major local sports clubs and preferred cuisine. For instance, the *giants* topic is prominent in San Francisco, highlighting the popularity of the local baseball team. Another characteristic example is the *pint* topic in London, which includes tokens such as *fish, chips, england* and *beer*.

C. Geographic Analysis of Topics

From the inspection of prevalent topics in major cities, we conclude that user-generated content in LBSNs exhibits trends that can be attributed to the rhythm of life typically observed in urban settings. Next, we will present a more fine-grained geographic analysis of topics that have emerged during our experiments for a particular city. In Figure 7(a) (at the end of the paper), we present the spatial distribution of a *cinema* topic (whose most probable words are *inception watching social story film watch network cinema scott pilgrim*) in central London. As highlighted by the intensity of the red squares on the map, this topic peaks in multiple neighbourhoods across the whole city at night, mostly in the immediate proximity of cinemas. In Figure 7(b), the *lake* topic (*boat lake beautiful water summer perfect michigan ride bike weather*) is a noteworthy case: Not only does this topic have a strong geographic focus, but it also follows nicely the coastline of the city of Chicago, showing how the dataset and topic modelling framework allow us to uncover geographic and temporal patterns within cities. We have observed similar cases for other cities as well.

Another interesting case is depicted in Figure 7(c), where one particular topic only shows up in the Borough of Brooklyn in New York City. This *Brooklyn* topic (*brooklyn williamsburg hipsters greenpoint vegan hipster hood northsidefest club games*) gives us insights into cultural trends in the area, such as the presence of a hipster community and the existence of vegan restaurants or clubs. Finally, we show a topic with top word *beautiful* (*beautiful gorgeous nice weather blue view walk beach angels festival*) detected in San Francisco (Figure 7(d)). It mainly covers parks, the coastline and even Treasure Island, and so gives a linguistic signature to areas particularly popular for leisure activities.

D. Geographic Entropy of Topics

As we have shown in the previous sections, topics may vary with respect to their semantic context, the time they become popular or their geographic coverage. We also observed that topics may have variable spatial shapes, i. e. they may cover a large number of areas or instead be restricted to specific neighbourhoods of a city. We use the standard information-theoretic entropy measure to capture the geographic dispersal of topics. In particular, we define the *geographic entropy* of a topic z as

$$E_z = -\sum_{l=1}^{L} \frac{c(w|l,z)}{c(w|z)} \log_2 \frac{c(w|l,z)}{c(w|z)}. \quad (4)$$

where $c(w|l,z)$ is the number of tokens assigned to topic z observed in area l and $c(w|z)$ is the total number of tokens assigned to the topic. We have measured the geographic entropy of topics across cities and present a histogram for each city in Figure 3. In the plots, small entropy values correspond to rather local topics restricted to only a few urban areas, while topics with higher entropy are spread across more neighbourhoods in the city. In our experiments, topics showed a wide range of entropy values, which implies a high variability in the geographic coverage of individual topics. An initial inspection suggests that topics in Chicago and New York are more entropic, whereas entropy values in London and San Francisco are more balanced.

An interesting question is what sorts of topics tend to be more or less entropic. In Table VI, we list the most and least geographically entropic topics inferred during our experiments. In general, low-entropy topics appear to relate to geographically focused and popular activities. This includes train stations (London's King's Cross), New York's Times Square (Figure 4(a)) and activity in popular venues such as San Francisco's Castro Theater (Table VI). On the other hand, topics with extensive geographic span corresponding to universal activities such as food, work etc. are often highly entropic. A visual inspection of two entropy extremes in New York is presented in Figure 4.

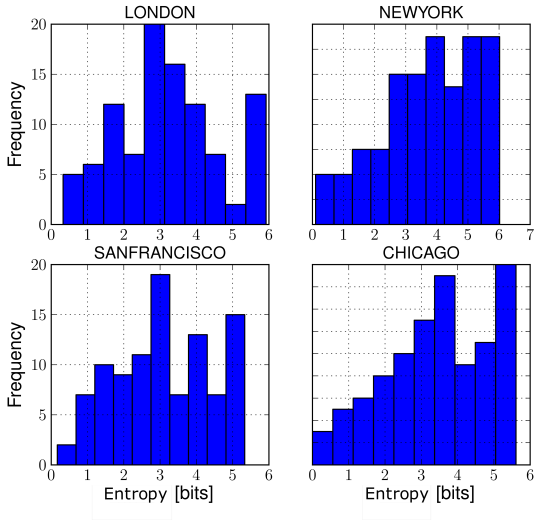


Fig. 3. Overall topic entropy in the four cities. On the y-axis we note the number of topics, and on the x-axis we measure entropy in bits.

E. Temporal Evolution of Topics

Next, we investigate whether the posterior topic distribution can also reveal patterns in the temporal domain. For this, we

chicago*	time	good	great	love	back	place	free	lets	long
party	happy	drink	drinks	stop	night	bday	house	spot	friends
birthday	tonight	event	sushi	friends	celebrating	wine	girls	finally	awesome
work	working	today	coffee	meeting	iced	time	days	pumpkin	weekend
dinner	drinks	show	food	pizza	date	place	beer	hour	wine
night	home	tonight	party	tomorrow	chillin	halloween	club	nite	karaoke
school	bike	change	pick	moving	move	session	road	church	college
morning	breakfast	coffee	today	friday	early	monday	start	week	work
food	burger	patio	chicken	burgers	post	perfect	fish	soup	chips
lunch	brunch	today	salad	sandwich	lunchtime	bloody	burrito	soup	special

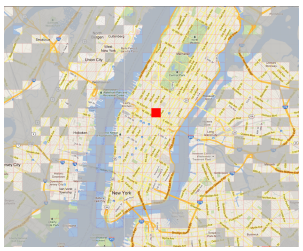
TABLE IV: Top 10 topics (in rows) inferred in Chicago. The background topic is marked with a star.

francisco*	good	great	back	love	time	awesome	place	finally	waiting
food	place	love	chicken	good	yummy	birthday	delicious	burger	favorite
dinner	night	birthday	party	tonight	happy	drinks	food	friends	pizza
time	work	today	break	game	afternoon	picking	hour	busy	weekend
happy	beer	hour	wine	halloween	late	friday	tasting	drinking	meal
coffee	work	working	early	meeting	office	latte	friday	check	lets
giants	game	lets	sfgiants	beat	world	series	baseball	opera	watching
meeting	stuff	store	black	shopping	found	bought	market	street	wrong
event	team	great	social	work	checking	meetup	meeting	conference	talk
home	back	heading	sweet	long	check	finally	train	leaving	headed

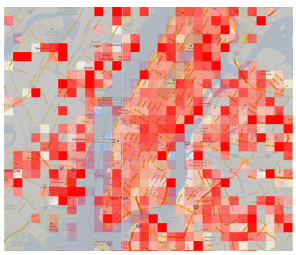
TABLE V: Top 10 topics (in rows) inferred in San Francisco. The background topic is marked with a star.

CHI	work	working	today	coffee	meeting	iced	time	days	pumpkin
LDN	pint	beer	lovely	pizza	chips	drink	england	fish	local
NYC	work	coffee	back	today	good	iced	office	working	latte
SF	coffee	work	working	early	meeting	office	latte	friday	check
CHI	cheezborger	wine	ontario	mowry	swine	directmail	sale	michigan	close
LDN	paris	eurostar	brussels	kings	north	leeds	harry	cambridge	pancras
NYC	park	bryant	square	broadway	times	movie	tourist	people	library
SF	castro	film	festival	hartford	dick	moby	badlands	smoothie	metropolis

TABLE VI: Most (top part with grey background) and least entropic Topics. We abbreviate Chicago (CHI), London (LDN), New York (NYC) and San Francisco (SF).



(a) NYC: topic *Times Square*



(b) NYC: topic *Work*

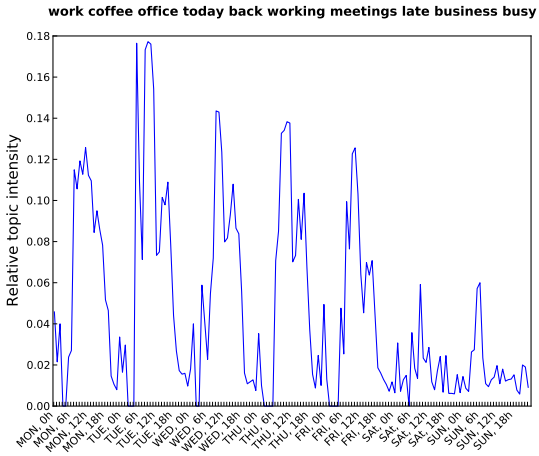
Fig. 4. Topics with extreme entropy values in New York. On the left, we show *Times Square* as the least entropic topic, whereas *Work* (right) is the most entropic one.

consider the total topic counts per time slot:

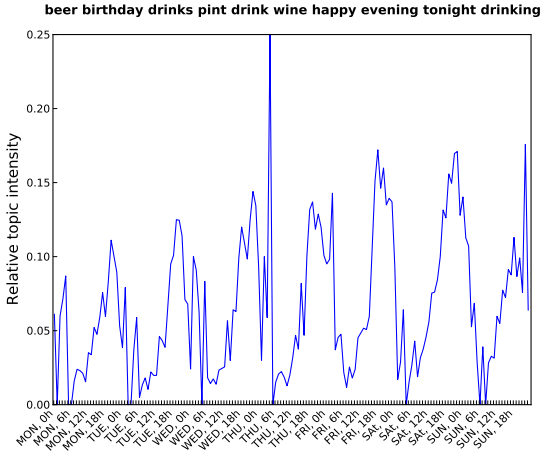
$$\text{intensity}_{t,z} = \frac{c(w|t,z)}{c(w|t)} \tag{5}$$

We have analysed the dataset using multiple topic model configurations, to reveal patterns of different temporal granularity. For example, if we use just the five time-of-day slots mentioned earlier, the model will detect prominent topics for *morning* and *evening*, but will not be able to determine that day-of-the-week patterns are present in the dataset as well. To mitigate this effect, we use 24 one-hour time slots per day of the week, and hence 168 in total. This way, we are able to represent both time-of-the-day- and day-specific trends, while at the same time keeping the time slots large enough to avoid sparsity problems.

Indeed the temporal evolution of prominent topics' intensities follows closely the rhythm of life of large cities in the industrialised world. We give some examples for London in Figure 5. For instance, work-related topics peak from Monday to Friday (cf. Figure 5(a)), while the values are much lower on weekends. There is an intra-day pattern as well: The topic's intensity is high only during normal office hours, except during lunch break when we expect much fewer meetings to take place. The same applies to the *party* topic (cf. Figure 5(b)), which regularly peaks in the evenings when people are most likely to go out with friends. What is less obvious, however, is that the topic's peak intensity also rises quite steadily from



(a) Work



(b) Party

Fig. 5. Temporal evolution of two topics in London. Note the most frequent words in the title.

day to day, until the weekend has arrived.

In order to capture the types of temporal topics that our model has detected, we treat the 168-dimensional vectors that describe the *intensity* of each topic over time as input points to a clustering algorithm. Our goal is to understand whether we can group the topics into meaningful clusters according to the temporal evolution of their intensities. We use a Spectral Clustering algorithm [16] to cluster the topics for New York City; using the eigengap heuristic [19], we set the number of clusters to 6. In Figure 6, we present the groupings of temporal topics that the clustering algorithm has yielded. To illustrate each of the six resulting clusters, we pick the topic with the highest number of tokens (maximum intensity) as a representative of the cluster. We notice that the classes cover a wide spectrum of weekly periodicities, with weekday and weekend topics being clearly separated. Similarly, night-time and dinner topics are discriminated from those that peak during lunchtime. On the top of each graph we present a list of words for the topic that intuitively corroborates the temporal aspects of each cluster.

F. Evaluation of topics with respect to place categories

We have so far analysed the geographic and temporal patterns exhibited by the inferred topics by aggregating all the textual content produced in a sub-area or time slice, but we have not taken into account further properties of the places. However, the dataset we collected from Foursquare also contains user-specified category names for most venues. For example, a venue might be classified as “Gym/Fitness”, “Hotel”, “Field” or “Train”. This allows us to examine whether certain topics are typical of places assigned to a particular category. In other words, we are able to characterise such categories of places by the most commonly discussed topics there, and thereby gain interesting insights into how people’s communication behaviour depends on the characteristics of the venue currently visited. In order to do this, we count, for each category, how often each topic was assigned to one of the places belonging to that category. We then sort the categories by the overall number of tokens produced at places belonging to each one of them. Results for five prominent categories of places in Chicago are given in Table VII. We can see that the inferred topics are highly typical of the categories in that people are very likely to produce content related to the category name. We have obtained similar results for the other cities as well.

G. Numerical Evaluation

Finally, we examine the predictive performance of our spatio-temporal topic model by computing the likelihood it assigns to unseen test data and comparing the resulting values to vanilla LDA as a baseline. The multi-partition structure of the dataset means that it is not straightforward to hold out a set of documents for testing. Instead, we randomly split the tokens into training and test sets; this is similar to the document completion evaluation used by [13]. We train the model on only a subset of all tokens, and then use the learned switching variable, topic and word distributions Π, θ and Ψ, as well as the re-estimated hyperparameters α and β, for predicting the likelihood of the held-out tokens. We use 90% of all tokens in the corpus for training and 10% for testing. Following the recommendations of [22], we use a so-called “left-to-right” procedure to estimate the log-probability assigned by each model to the test set. We average the results of 20 runs (or “particles”) on the same test set to obtain the final result. In Table VIII, we report the total log-likelihood divided by the number of tokens in the test set. We can see that our model assigns higher likelihood to the test data than standard LDA for all four cities considered, although the differences are small in each case. We leave it to future work to investigate the advantages of our spatio-temporal topic model over vanilla LDA.

V. Related Work

Our work brings together and is related to two different fields, namely the area of Location-based Social Networks and that of Topic Modelling and Natural Language Processing. Despite Location-based Social Networks being a very recent class of service, there is already a large body of work which addresses a variety of problems. For example, the relationship

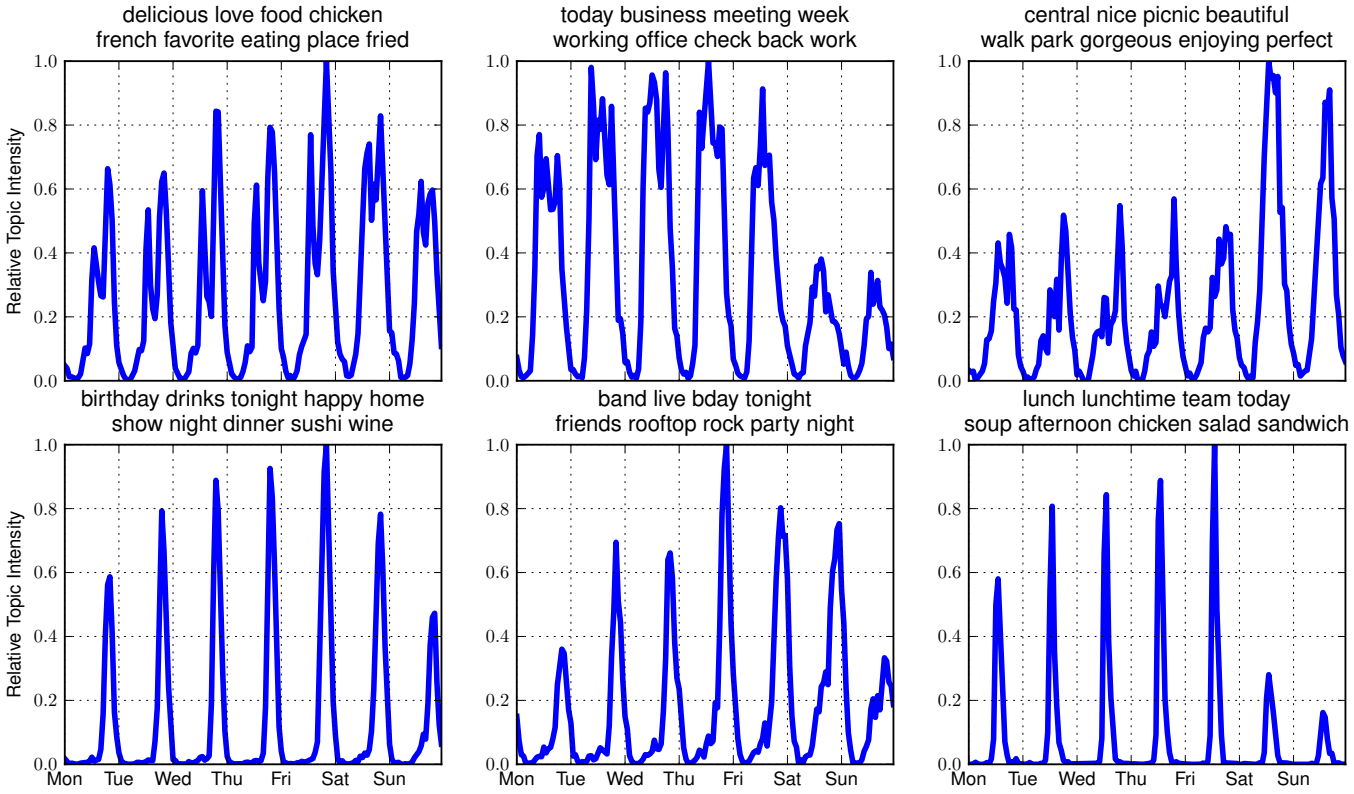


Fig. 6. Representative topics of the six clusters that have emerged after clustering the temporal vectors of topics in New York. We can observe qualitatively different types of temporal topics including the *weekend* topic (top right) and the *dinner* and *lunch* topics (bottom left and right).

City	Spatio-temporal	Baseline
London	-8.22	-8.29
New York	-8.43	-8.44
Chicago	-8.16	-8.20
San Francisco	-8.27	-8.35

TABLE VIII: Log-likelihood assigned to test tokens (averaged, per token).

between mobility and friendships in LBSN has been studied by Cho et al. [4], while in [14] Scellato et al. investigate the spatial properties of such networks. Research on potential applications includes that of link prediction in Location-based Services [15] and the prediction of the categories of Foursquare places based on the temporal check-in patterns of visiting users [25]. A few studies have applied topic models such as LDA to non-linguistic datasets arising from location-based services such as sets of location category tags [5] and time-location coordinates [7]; these studies do not use the content of messages at all. A distinguishing characteristic of data from Foursquare and similar LBSNs is that it pairs linguistic content with spatio-temporal metadata; our study illustrates the potential for research which considers these dimensions jointly.

Since Blei et. al. introduced Latent Dirichlet Allocation (LDA) as a generative probabilistic model for text corpora [2], numerous extensions have been proposed which integrate various kinds of metadata available alongside documents. Some of this work has focused on the temporal dynamics of timestamped documents, either by modelling how topics change over time [1], [23], or by generating timestamps in a similar fashion to words [24]. More recently, an alternative approach [9] to capturing temporal topic evolution in a large corpus was proposed which clusters topics into semantically coherent subgraphs and arranges them on a timeline. In general, this strand of research has focused on modelling corpora of scientific or news articles rather than social media content. There is also an increasing interest in topic models that incorporate geographic information. Both [6] and [26] aim to discover areas (of a country) that are coherent in both space and language use; the former learns models of Twitter data in order to find regionally specific vocabularies, while the latter models the spatial distribution of photograph tags on Flickr.

Our model differs somewhat from most of the aforementioned spatial and temporal models in that we do not use these dimensions to regularise our topic model; we do not assume that the language used in early and late mornings will be similar, for example. The impact of these modelling choices is one that we intend to investigate in future work. The most similar approach to ours is that of [10], whose model shares a number of structural features, including the use of a background component independent of spatial and temporal influences. One perspective on our model is that it is a Bayesian analogue of Mei et al.'s EM-trained model [10], with the robustness advantages that are brought by the use of priors. In this sense, it is a natural progression with respect to the previous literature: for the first time both temporal and geographic variables are combined in the context of a Bayesian topic model to infer topics from a spatio-temporally annotated

Category	Topic top words								
Hotel	chicago	hotel	view	nice	floor	river	tour	beautiful	downtown
	work	today	meeting	working	coffee	back	office	days	business
	party	night	drinks	stop	drink	tonight	show	late	friends
Train	home	back	train	heading	headed	waiting	late	line	downtown
	work	today	meeting	working	coffee	back	office	days	business
	stop	favorite	late	hair	break	shop	quick	afternoon	cool
Field	softball	game	team	tennis	park	soccer	league	playoffs	playing
	work	today	meeting	working	coffee	back	office	days	business
	stop	favorite	late	hair	break	shop	quick	afternoon	cool
Gym/Fitness	workout	time	fitness	working	work	cardio	yoga	sexy	body
	work	today	meeting	working	coffee	back	office	days	business
	morning	coffee	breakfast	work	early	today	start	week	ready
Fast Food	food	place	love	yummy	good	cheese	delicious	chicken	hungry
	lunch	salad	lunchtime	sandwich	chicken	burger	soup	eating	pizza
	work	today	meeting	working	coffee	back	office	days	business

TABLE VII: Most frequent place categories in Chicago along with the three most prominent topics (one per line) at places belonging to that category.

corpus.

More generally, researchers in text mining and computational science have found social media to be a very valuable resource for large-scale investigations of human behaviour, linguistic and otherwise. Space restrictions do not permit us to provide a comprehensive overview; sample topics of investigation include dialogue dynamics [12], user profiling [11] and daily mood patterns across the globe [8].

VI. Conclusions

In this work, we have investigated the potential of using topic modelling techniques for detecting trends in geo-tagged microblog posts, using a dataset of millions of Foursquare comments. We have also introduced a novel spatio-temporal topic model based on the use of a switching variable, which explicitly takes into account the geo-temporal side information on a per-token basis. We have demonstrated that the abundance of data available in LBSN enables such models to capture the topical dynamics in urban environments.

We expect the present work to be a useful tool for social scientists who may use it to examine cultural trends in urban environments which are reflected in people's communication behaviour. Also, developers of recommendation systems might consider exploiting linguistic features in order to make their recommendations more precise. In addition, content delivery in mobile applications and advertising content could be tuned to fit the local and temporal trends of topics emerging in Location-based Social Networks.

References

[1] D. M. Blei and J. D. Lafferty. Dynamic Topic Models. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 113–120, New York, NY, USA, 2006. ACM.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[3] C. Chemudugunta, P. Smyth, and M. Steyvers. Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model. In *Proceedings of NIPS '06*, Vancouver, BC, 2006.

[4] E. Cho, S. A. Myers, and J. Leskovec. Friendship and Mobility: User Movement In Location-Based Social Networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1082–1090, New York, NY, USA, 2011. ACM.

[5] J. Cranshaw and T. Yano. Seeing a home away from the home: Distilling proto-neighborhoods from incidental data with Latent Topic Modeling. In *Proceedings of the NIPS Workshop on Computational Social Science and the Wisdom of Crowds*, 2010.

[6] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A Latent Variable Model for Geographic Lexical Variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1277–1287, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[7] L. Ferrari, A. Rosi, M. Mamei, and F. Zambonelli. Extracting Urban Patterns from Location-based Social Networks. In *Proceedings of LBSN '11*, Chicago, IL, 2011.

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

[9] Y. Jo, J. E. Hopcroft, and C. Lagoze. The Web of Topics: Discovering the Topology of Topic Evolution in a Corpus. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 257–266, New York, NY, USA, 2011. ACM.

[10] Q. Mei, C. Liu, H. Su, and C. Zhai. A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 533–542, 2006.

[11] M. Pennachiotti and A.-M. Popescu. Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter. In *Proceedings of KDD '11*, San Diego, CA, 2011.

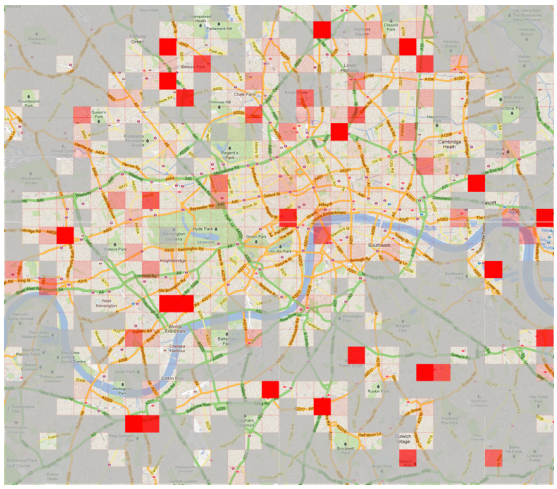
[12] A. Ritter, C. Cherry, and B. Dolan. Unsupervised Modeling of Twitter Conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 172–180, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[13] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The Author-Topic Model for Authors and Documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 487–494, 2004.

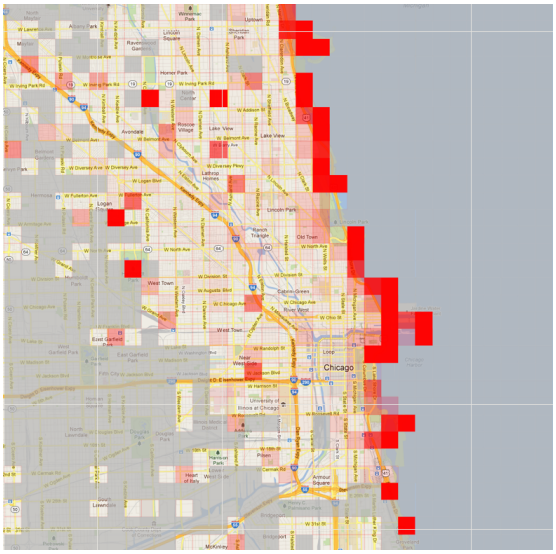
[14] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial Properties of Online Location-based Social Networks. In *Proceedings of ICWSM' 11*, July 2011.

[15] S. Scellato, A. Noulas, and C. Mascolo. Exploiting Place Features in Link Prediction on Location-based Social Networks. In *Proceedings of KDD '11*, New York, NY, USA, 2011. ACM.

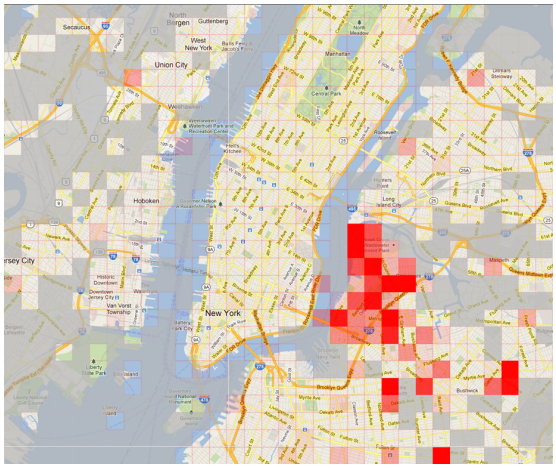
[16] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, page 731, Washington, DC, USA, 1997. IEEE Computer Society.



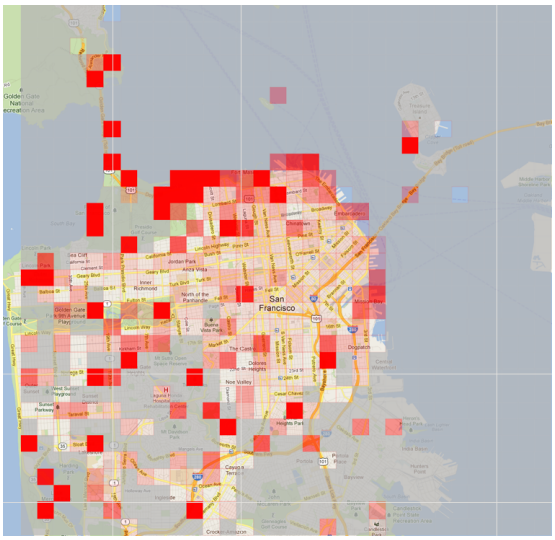
(a) London: *Cinema* topic



(b) Chicago: *Lake* topic



(c) New York: *Brooklyn* topic



(d) San Francisco: *Beautiful* topic

Fig. 7. Selected topics for each city. The intensity of the red color in each square reflects the intensity value of the presented topic at that area of the city. With grey we depict all areas for which there is no data and have been discarded.

[17] Techcrunch. Foursquare now 3 million strong. http://techcrunch.com/2010/08/29/foursquare-now-3-million-strong/, 2010.

[18] The Next Web. Foursquare hits 20 millions users and 2 billion check-ins. http://thenextweb.com/socialmedia/2012/04/16/foursquare-hits-20-million-users, April 2012.

[19] U. von Luxburg. A Tutorial on Spectral Clustering. Technical Report 149, Max Planck Institute for Biological Cybernetics, August 2006.

[20] H. Wallach, D. Mimno, and A. McCallum. Rethinking LDA: Why Priors Matter. In *NIPS*, 2009.

[21] H. M. Wallach. *Structured Topic Models for Language*. PhD thesis, University of Cambridge, 2008.

[22] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation Methods for Topic Models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1105–1112, New York, NY, USA, 2009. ACM.

[23] C. Wang, D. Blei, and D. Heckerman. Continuous Time Dynamic Topic Models. In *Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence*, pages 579–586, 2008.

[24] X. Wang and A. McCallum. Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 424–433, New York, NY, USA, 2006. ACM.

[25] M. Ye, D. Shou, W.-C. Lee, P. Yin, and K. Janowicz. On the Semantic Annotation of Places in Location-Based Social Networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 520–528, New York, NY, USA, 2011. ACM.

[26] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang. Geographical Topic Discovery and Comparison. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 247–256, New York, NY, USA, 2011. ACM.