LoCaTe: Influence Quantification for Location Promotion in Location-based Social Networks

Ankita Likhyani
Indraprastha Institute of Information Technology, Delhi, India
ankital@iiitd.ac.in

Srikanta Bedathur
IBM-Research Lab, Delhi, India
sbedathur@in.ibm.com

Deepak P.
Queen's University Belfast Northern Ireland, UK
deepaksp@acm.org

Abstract

Location-based social networks (LBSNs) such as Foursquare offer a platform for users to share and be aware of each other's physical movements. As a result of such a sharing of *check-in* information with each other, users can be influenced to visit at the locations visited by their friends. Quantifying such influences in these LBSNs is useful in various settings such as location promotion, personalized recommendations, mobility pattern prediction etc. In this paper, we focus on the problem of location promotion and develop a model to quantify the influence specific to a location between a pair of users. Specifically, we develop a joint model called *LoCaTe*, consisting of (i) user mobility model estimated using kernel density estimates; (ii) a model of the semantics of the location using topic models; and (iii) a model of time-gap between check-ins using exponential distribution. We validate our model on a long-term crawl of Foursquare data collected between Jan 2015 – Feb 2016, as well as on publicly available LBSN datasets. Our experiments demonstrate that LoCaTe significantly outperforms state-of-the-art models for the same task.

1 Introduction

With GPS[1] becoming pervasive within smartphones, location-based social networks (LBSNs) such as *Foursquare* and *Facebook places* have gained popularity. These networks allow users to share their check-in information with their friends enabling LBSN users to be aware of their friend's check-ins. This, combined with other factors, can influence the check-in preferences of users [Cho *et al.*, 2011; Ye *et al.*, 2011]. Quantifying the user-to-user influence —often modelled probabilistically [Goyal *et al.*, 2010; Zhang *et al.*, 2012]— is an essential building block in understanding the effects of information diffusion in LBSNs.

Influence maximization [Bouros *et al.*, 2014; Li *et al.*, 2014; Wu and Yeh, 2013] in social networks addresses the problem of finding a set of users who have a strong influence in the social network; these users are potentially good seed users to run promotion campaigns that target to maximize the reach of the campaign. Research interest in LBSNs has sparked interest in using geo-locations for influence maximization, leading to the *location promotion* problem [Zhu *et al.*, 2015]. The location promotion problem is of significant importance for launching effective campaigns to help businesses gather more customers. This task instantiates the influence maximization task on a specified target location/venue (e.g., a particular restaurant), with the intent of finding a set of seed users who are well-positioned for the promotion of their business [Zhu *et al.*, 2015]. Once a set of seed users is identified, it can be used to issue targeted special offers.

From the perspective of location promotion, the *worth* of a user - or the user's "seedness" - is quantified as the number of users she can influence to visit a location, after her visit. The necessary step for the location promotion problem is thus a method to assess the likelihood that other users would visit the location following the candidate seed user's visit. Influence quantification can factor in a variety of information that a LBSN offers: (i) geographic features: user's mobility over different locations, (ii) semantic features: type/category of location (e.g., restaurant, cafe), (iii) social correlation: the relationship between users in the social network, and (iv) temporal correlation: the degree to which a user's movement is correlated with another user's. Previous work on influence quantification for location promotion has mostly focused on modelling geographic features and social correlation [Zhu *et al.*, 2015]. Studies on semantic features such as category have been limited primarily since datasets containing such information have been scarce [Gao *et al.*, 2012; Cho *et al.*, 2011]; such deficiencies have been recently addressed by methods such as approximate spatial joins [Likhyani *et al.*, 2015]. Thus, because of unavailability of category information intrinsically, the use of location semantics have not been explored in previous work. In this work, we join datasets from multiple sources to make location category usage possible.

The temporal correlation of users behavior has been modeled previously in online social networks, but not in LBSN as we do in this paper. The socially induced followship based on temporal correlation has been of interest in LBSN studies in other contexts [Pham and Shahabi, 2016].

In this paper, we address this research gap and make twofold contribution: first, exploitation of the hitherto unexplored information, that of location categories, for influence quantification; second, we develop a novel model **LoCaTe**

[1] https://en.wikipedia.org/wiki/Global_Positioning_System

that incorporates:

1. **Lo**cation affinity: The mobility patterns of users that hold cues to whether they frequent the proximity of the target location.
2. **Ca**tegory affinity: The affinity of a user to the *category* or type of the location.
3. **Te**mporal correlation: The temporal correlation of movements between the user and the candidate seedset, thus modelling time-conditioned social followship.

As is the case of any influence quantification method, our model easily fits into any location promotion setting[Li *et al.*, 2014; Zhu *et al.*, 2015]. We empirically evaluate our approach over both publicly available real-world LBSN data as well as our own long-term crawl of Foursquare check-ins spanning more than one year. Overall, the proposed LoCaTe model can outperform state-of-the-art methods by more than 50% in AUC (Area Under the Curve of ROC[Bradley, 1997]).

2 Related Work

While influence maximization has been a well-studied problem (e.g., [Kempe *et al.*, 2003; 2005; Chen *et al.*, 2010; Goyal *et al.*, 2010]), the geo-seeded instantiation motivated by LBSNs has gathered only recent attention [Li *et al.*, 2014; Zhu *et al.*, 2015; Bouros *et al.*, 2014; Zhang *et al.*, 2012; Pham and Shahabi, 2016]. Apart from the location promotion problem where we start with a specific target location, there have been studies on *region promotion*, where the target is a larger geo-region [Bouros *et al.*, 2014]. Users' geo-location affinities have been modeled by either associating one specific geo-location with each user (usually the most frequently one visited by the user) [Li *et al.*, 2014; Zhang *et al.*, 2012] or a set of geolocations or only the social network structure [Bouros *et al.*, 2014; Zhu *et al.*, 2015]. In a similar way, the user-user pairwise influence propagation probabilities are estimated either using just the (social) network structure [Bouros *et al.*, 2014; Li *et al.*, 2014; Zhang *et al.*, 2012] or taking into consideration the seed location/region [Zhu *et al.*, 2015]. To the best of our knowledge only the recent work in [Pham and Shahabi, 2016], have looked at defining user-user pairwise influence in spatio-temporal context, but for identifying followship. For our empirical evaluation, we compare against the most recent work by Zhu et al. [Zhu *et al.*, 2015], that associates a set of locations for each user and considers the influence between two users to be dependent on the location.

User Mobility Models: We make use of user mobility models in our method drawing inspiration from earlier work on characterizing user behavior in LBSNs. Since LBSN data provides a trail of user's locations, it provides a rich data platform for studying user mobility patterns; such patterns are of interest for tasks such as *location prediction* and *personalized recommendation*. In literature, mobility models that mine spatial patterns based on generative models [Gao *et al.*, 2012], Gaussian distributions [Cho *et al.*, 2011] and kernel density estimations [Lichman and Smyth, 2014] have been particularly successful. Accordingly, we use the kernel density mobility model [Lichman and Smyth, 2014] to model and exploit user-location affinities in our method for LBSN location promotion. The baseline technique from [Zhu *et al.*, 2015], on the other hand, makes use of a distance-based mobility model, DMM, in their influence quantification method.

Symbol	Description
G	A location based social network
U	Set of users in G
E	Set of connections from u_i to u_j s.t. $u_i, u_j \in U$ and $u_i \neq u_j$
ℓ	A location specified by a triple (x, y, C_ℓ), where x,y correspond to geo-coordinates and C_ℓ to category set of ℓ
$\langle u, \ell, t, C_\ell \rangle$	A check-in record of user u at time t at location ℓ that has a category set C_ℓ
M_u	set of check-in records $\langle u, \ell, t, C_\ell \rangle$
L	A set of locations
C	A set of categories

Table 1: Notations used in this paper

3 Problem Statement

Now we provide a formal definition of influence quantification problem in an LBSN for location promotion. Table 1 lists a set of notations that will be used. We model a location/venue[2] as having a fixed geographic coordinate as well as a set of categories associated with it. This allows for modeling of locations such as movie multiplexes that would screen movies as well as contain eateries.

Definition 1 (Location Promotion) *Given an LBSN G, a target location ℓ, whose category set is C_ℓ, the location promotion problem is to select a small set of seed users S, $S \subseteq U$, who can lure other users to the target location ℓ well. The hyperparameter τ indicates the desired cardinality of S.*

Influence quantification, the task of quantifying the likelihood of a user visiting the location given a visit to the same location by another (i.e., seed) user, forms an important building block for location promotion problem as shown in [Zhu *et al.*, 2015]. Given a user-user influence model, the location promotion problem reduces to finding a set of seed users which can collectively influence a large set of users to visit the location in focus.

Definition 2 (Influence quantification.) *Influence quantification: Given an LBSN G, a target location ℓ, a seed-user u (usually a user who has previously visited ℓ), the influence quantification problem is to quantify the likelihood $P(\ell, u, v \mid G)$, the likelihood that any user v among u's connections is likely to visit ℓ.*

As mentioned earlier, influence quantification models typically consider the affinity of the user v to location ℓ independent of the influence from other users, and the influence of the seed user u on v in this decision.

Evaluation: Our evaluation framework follows earlier work in this area (e.g., [Zhu *et al.*, 2015]): from a dataset of LBSN check-ins, we decide on a cut-off timestamp such that there is a sizeable amount of check-ins before and after the cut-off. All data prior to the cut-off timestamp is used as training data to learn the model. The remaining data forms the test set against which the effectiveness of the learned model is measured. Consider a particular instance of the influence quantification problem for location ℓ and a seed-user u; the

[2]We use location and venue interchangeably. Though *venue* might be more appropriate, *location* is also used in literature.

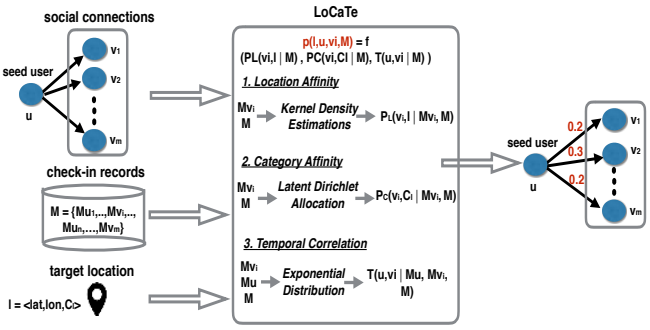


Figure 1: LoCaTe Framework for Influence Quantification

influence quantification output would be an ordered list of u's connections, ordered in the decreasing (non-increasing) order of estimated likelihood to visit ℓ. This list can be cut-off using a threshold ρ to identify a set of users who are deemed to be highly likely to visit ℓ - this set forms the *predicted set*, $PS(\ell, u, \rho \mid G)$. The ground truth activated set, $A(\ell, u)$, is the subset of u's connections who have actually visited ℓ after the cut-off timestamp (i.e., from the test set). The match between $PS(\ell, u, \rho \mid G)$ and $A(\ell, u)$ measured at various values of the threshold ρ quantifies the goodness of the influence quantification method employed. Any measure of match between sets can be aggregated over all users (i.e., by iterating u over the set of LBSN users) to get a single goodness value for the combination $[\ell, \rho]$. We will use the ROC curve (generated by varying ρ) to compare our method against baselines in our empirical evaluation.

4 LoCaTe Influence Quantification

We now outline our influence quantification method, LoCaTe, that estimates $P_{\ell,u}(v|M)$, a scoring that captures the likelihood that the user v from u's connections would visit the location ℓ using the check-in records in the training part, denoted as M. Figure 1 shows the framework of *LoCaTe*. LoCaTe combines information from three kinds of features:

$$P_{\ell,u}(v|M) = \Big(\alpha \underbrace{P_L(v,\ell|M)}_{\text{location affinity}} + (1-\alpha) \underbrace{P_C(v,C_\ell|M)}_{\text{category affinity}}\Big) \times \underbrace{T(u \to v|M)}_{\text{temporal correlation}},$$

where, $P_L(v,\ell|M)$ models the affinity of v to location ℓ, and $P_C(v,C_\ell|M)$ models the affinity of v to the categories that are associated with the location ℓ (denoted as C_ℓ). These two terms are interpolated using an interpolation parameter α. Further, $T(u \to v|M)$ captures the temporal correlation between users u and v, *independent* of the location. The first two terms quantify users affinity for the location using mobility and categories respectively and are combined using a weighted sum. The third term quantifying location-agnostic user-user socially induced temporal affinity is merged using a product. Thus, the final scoring ensures that users who are strong on both location and temporal aspects score much higher than others.

$P_{\ell,u}(v|M)$ ranges between [0,1], since we normalize the scores obtained for a list of influenced users $v = (v_1, ...v_m)$ keeping the seed user u and target location ℓ fixed. Thus

$$\Sigma_v P_{\ell,u}(v|M) = 1.$$

The usage of **Lo**cation affinity, **Ca**tegory affinity and **Te**mporal correlation lends the name to our method.

4.1 Location Affinity

The mobility of each user is typically restricted to a few key locations, which would typically include the location of stay and work [Cho *et al.*, 2011]. Thus, a user has an inherent preference for some geo-locations. We use kernel density estimates to model mobility patterns and location affinity of users. Kernel density estimates are robust to sharp transitions in spatial densities that human mobility witnesses, especially in contexts involving travels that take users far away from their usual location of residence [Cho *et al.*, 2011; Lichman and Smyth, 2014].

The affinity of v to ℓ based on her own check-in history is modeled as the kernel density estimate (KDE) that quantifies the average weighted similarity between ℓ and each location l_j has checked-in using hyper-parameter k,

$$P_L(v,\ell|M_v) = f_{KD}(l|M_v,k) = \frac{1}{|M_v|}\Sigma_{j=1}^{|M_v|} K_{j,k}(\ell,\ell_j)$$

$K_{j,k}(\cdot,\cdot)$ estimates the similarity between locations as inversely related to the Euclidean distance between them:

$$K_{j,k}(\ell,\ell_j) = \frac{1}{2\pi h_{j,k}} \exp\left(-\frac{1}{h_{j,k}}\|\ell - \ell_j\|\right)$$

Here, $h_{j,k}$ is a location-dependent scalar factor that is set to be the Euclidean distance of ℓ_j to it's k^{th} nearest neighbor, thus ensuring that the similarity computation is sensitive to differential densities of locations in urban and rural areas.

Mixture of Kernel Density models:

To offset for sparsity issues in determining the location affinity using just the user's check-in records, we use corpus smoothing by interpolating the location affinity of a user with those across all users, yielding our final formulation:

$$P_L^k(v,\ell \mid M) = \beta_v f_{KD}(\ell \mid M_v, k) + (1-\beta_v) f_{KD}(\ell \mid M, k),$$

where, β_v is a user-specific mixing weight, determining the relative influence between the user and global models. We will denote this as $P_L(.,.)$ when the value of k is clear.

Training:

We have two parameters that need to be estimated. First, the hyper-parameter k is estimated as the value that maximizes the likelihood of check-ins in a chosen validation set, V (which is a held-out part of the check-ins before the cut-off timestamp, but not included in the training). Thus, we set k to the value that maximizes the following:

$$k = \arg\max_{k'} \Sigma_{\langle v,\ell,\cdot,\cdot\rangle \in V} \log\left(P_L^{k'}(v,\ell \mid M)\right)$$

The distribution of log-likelihood across various values of k are shown in Table 2; accordingly, we chose $k = 5$ for usage in our method. Second, we choose the user-specific interpolation weights β_v as the value that maximizes the likelihood of their check-ins in the training data itself. We do not use the validation set here since there are users who do not have any check-ins in the validation set.

k	2	3	4	**5**	6	7	8	9	10
Fsq'16	-2.032	-1.804	-1.704	**-1.640**	-1.670	-1.687	-1.722	-1.744	-1.817
Fsq'11	-2.711	-2.640	-2.063	-1.726	-0.939	-0.738	**-0.677**	-0.794	-0.851
Fsq'10	-1.283	-1.251	-1.233	**-1.211**	-1.225	-1.231	-1.246	-1.260	-1.278
Brightkite	-1.915	-1.869	-1.836	-1.789	**-1.779**	-1.821	-1.850	-1.879	-1.897
Gowalla	-1.978	-1.896	-1.847	**-1.804**	-1.825	-1.854	-1.877	-1.890	-1.931

Table 2: Log-likelihood at different values of k

(a) User A (b) User B

Figure 2: Category wise check-in Distribution

4.2 Category Affinity

Locations/Venues often record correlated check-in behavior across LBSN users. For example, a restaurant might be better off targeting a user who frequently checks in to food places due to the correlation across various categories of food joints. As an example, consider two users in Figure 2 represented by the word cloud of the categories of their checked-in locations (larger font indicates higher frequency); User A evidently exhibits affinity towards visiting restaurants while user B prefers gym and fitness centers. We use topic modeling to identify such higher-level contexts, and exploit it to model the user-category affinity term, $P_C(v, C_\ell|M)$.

Latent Dirichlet allocation [Blei *et al.*, 2003] models semantic matching between text documents by learning topics, each of which is a probability distribution over the set of words. LDA ensures that words that are semantically related would have high probabilities associated with the same topic(s). In our adaptation of LDA for modeling topical contexts across check-in categories, each user v is treated as a document constructed as a bag of categories v_C (i.e., each category as a word) of checked-in locations. These documents across the users in the population form a document corpus. We apply LDA on this document corpus, to learn topics which are probability distributions over categories. We then use the learnt topics to estimate the user's affinity to the set of categories associated with the location of interest:

$$P_C(v, C_\ell|M) = \Sigma_{Z \in Topics(M)} P(C_\ell|Z) \times P(v|Z),$$

where $Topics(M)$ is the set of topics learned as described, and Z represents a topic from the topic-set learned. $P(v|Z)$ and $P(C_\ell|Z)$ quantify how well the category distribution for Z matches against those of the check-ins of v and the categories of ℓ respectively. High values of $P_C(v, C_\ell|M)$ are achieved when the user's category distribution and that of the location under consideration are correlated with the same set of topics.

4.3 Temporal User Correlation

We now turn our attention to the temporal correlation term, $T(u \rightarrow v|G)$, that quantifies the extent of influence that u has over v. This primarily accounts for the socially induced followship in our Influence Quantification model. The task at hand is to quantify the chance that v will follow u in checking-in to a location, such that $(u, v) \in E$. Towards

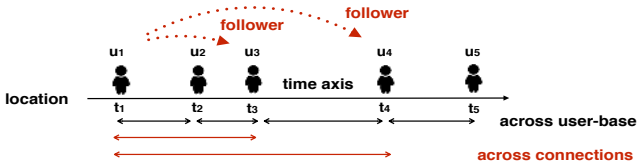
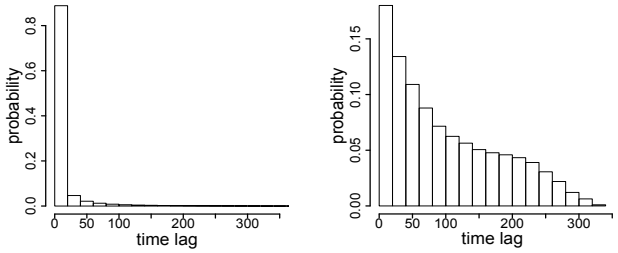


Figure 3: Depicting the time lag between check-ins at a location for connected and non-connected users



(a) across userbase (b) across connections

Figure 4: Time lag probability distribution plot

modeling this, we first analyze the behavior of general inter-arrival times of users in the LBSN at a given location, without considering whether they are connected to each other in the LBSN network or not; we call this the *time lag* distribution *across userbase*. The analogous time lag distribution *across connections* considers the distribution of the time duration elapsed between two users who are connected to each other, visiting the location in question. A visual depiction of the time lags that go into either of these distributions appears in Figure 3. We collect these time lag distributions across all locations in the LBSN and analyze their frequency distribution using a histogram. As expected, the general *across userbase* time lag distribution follows a power law distribution, as illustrated in Figure 4(a). However, the *across connections* time lag distribution in 4(b) does not quite follow a power law distribution despite exhibiting a monotonic decay with increasing values of time lag. It may also be noted that the across connections data is much sparser than across userbase; this is so since there are a significantly fewer number of occurrences of connected users visiting the same location. Drawing cues from the trends across the two time lag distributions, we choose to model the time lag distribution between users using an exponential distribution, given it's popularity in similar contexts [Pham and Shahabi, 2016].

Modeling using exponential distribution

According to the exponential distribution modelling, the weight associated with any value of time lag, denoted δt, would be quantified as the following:

$$p(\delta t) = \lambda_t e^{-\lambda_t \delta t}$$

We set λ_t to be the mean time lag between check-ins by connected users:

$$\lambda_t^{-1} = avg\{|t_2 - t_1| \mid \exists\langle u, \cdot, t_1, \cdot\rangle \in M \wedge \exists\langle v, \cdot, t_2, \cdot\rangle \in M \wedge (u, v) \in E\}$$

where the $\langle u, \cdot, t, \cdot\rangle$ implies that we consider all check-ins by u at time t irrespective of the location of the check-in or the set of categories associated with the location. This feeds into our user correlation estimate $T(u \rightarrow v|G)$ which is modelled

as the cumulative weight of v checking in at a location visited by u after a time lag of any $t \geqslant t_{u,v}^{min}$:

$$\begin{aligned} T(u \to v|G) &= \int_{t_{u,v}^{min}}^{\infty} \lambda_t e^{-\lambda_t \delta t} d(\delta t) \\ &= -e^{-\infty} + e^{-\lambda_t t_0} = e^{-\lambda_t t_{u,v}^{min}} \end{aligned}$$

$$t_{u,v}^{min} = min\{(t_2 - t_1) \mid \exists\langle u, \cdot, t_1, \cdot\rangle \in M \wedge \exists\langle v, \cdot, t_2, \cdot\rangle \in M\}$$

As indicated above, we set $t_{u,v}^{min}$ to be the earliest time that v has checked in after u at the same location, according to training data; this ensuring that $T(u \to v|G)$ reflects the extent of correlation between u and v, since $T(u \to v|G)$ would have a high value for those user pairs where the latter follows the former (temporally) closely.

5 Experiments

5.1 Experimental Setting

We tested over 5 datasets as shown in table 3, of which FSq'16 is the one that we collected using Twitter and FourSquare APIs, and rest are publicly available datasets [Cho *et al.*, 2011; Gao *et al.*, 2012]. The publicly available datasets do not provide category information, and moreover, the locations are also annonymized, they only expose the latitude, longitude information of the location. A recent work uses approximate spatial joins across checked-in locations and map data (available at Foursquare itself) to infer category information for each check-in and the data is made available by the authors [Likhyani *et al.*, 2015]. Such joins provide a large set of categories for each location as there could be multiple types of locations (e.g., cafe, hospital) at the same geo-location. Although another large collection of Foursquare check-in data with category information is released, it does not have the social graph of users essential for the location promotion problem addressed in this paper [Yang *et al.*, 2016]. To the best of our knowledge, there does not exist any publicly available dataset that contain category information of checked-in locations along with the social graph.

Data Collection: We use Twitter for getting check-ins of users, because user's check-in information in Foursquare can only be accessed from her own social circle, and it is not available publicly. However, Foursquare users can choose to post their check-ins via Twitter when they check in at a place. Hence, we capture check-ins by crawling Foursquare-tagged tweets from the Twitter Public Stream API [3] and REST API[4]. The Foursquare-tagged check-in tweet crawled contains a uniqueURL that points to a Foursquare web page containing the geographical information of this check-in location. We parse this web page to get check-in location along with the timestamp. Further to obtain category of this location, we pass this location to FourSquare API[5] and get access of the location information. Lastly, to accquire the social information of users (i.e. friendship links), we again use Twitter, because Foursquare does not provide public access to user's friends list. Thus, we acquired friendship ties that Foursquare users have among them on Twitter, where they are publicly available. To extract the check-in details of friends we again crawl their tweets in the same manner as above. While the resulting social graph is not identical to the original Foursquare graph, it encompasses the subset of users who are on Twitter and have linked their Twitter and FourSquare accounts. In our experiment, we consider the users who have at least 10 check-ins. Some key statistics of the dataset are shown in table 3. FSq'16 is the dataset that we curated, rest are public datasets for which we collected only the category information.

Test Set: For each dataset, we assign a cut-off timestamp, the data prior to it is used for training and rest of the check-ins for testing. The cut-off timestamp is chosen such that 80% of total checkins are used for training. The target locations are identified as locations where a user's check-in precedes its follower's check-ins and both the check-ins are made on or after the cut-off timestamp. The user who checks-in first is the *seed user* and it's followers who checks-in after him, are the $activated\ users$ (by the seed user). Section 2 describes ground truth construction set of activated users and also the evaluation. Table 3 shows the number of test cases ($A(l, u)$), and cut-off timestamp for each dataset. Note that there is a high variability in the number of categories per location across the datasets, thus enabling the empirical evaluation to cover scenarios across a wide spectrum of semantic information availability.

Parameter Estimation: We also construct a validation set in the same manner as test set is built from the training set for learning the parameters β_v and α : mixing the kernel density estimations for spatial density and the topic model for category information, respectively. We use EM algorithm to learn this. Table 4 shows values of β_v and α : learned for different datasets. The parameter Z (number of topics) is set to 50.

Implementation: We coded all the models including the baselines (DMM_Basic and DMM_Social), in Java. For KDE we made use of the source code available at UCI Datalab website (`http://www.datalab.uci.edu/resources`) and for LDA based topic model we made use of the source code available at `http://mallet.cs.umass.edu/topics-devel.php`. We ran the code on a 6-core 2.5GHz Intel Xeon CPU with 64GB of RAM. The source code and the datasets used are available publicly [6].

5.2 Comparative Evaluation

We compare our proposed model **LoCaTe** with two state of the art users' mobility based influence quantification models, *viz.*, the DMM_Basic and the DMM_Social models proposed in [Zhu *et al.*, 2015] and two baseline methods: GMM (Gaussian Mixture Model) and a baseline method that plugs in mobility, categorical and temporal features in a simplistic manner (described as follows).

DMM_Basic: It models the probability of a user moving from visited locations to the target location. Pareto distribution [Newman, 2005] is used for modeling the distances between the checked-in locations of a user.

$$\begin{aligned} p_u(\ell) &= \Sigma_l P(u \text{ is at } l)\ P(u \text{ moves distance } d(l, \ell) \text{ from } l) \\ &= \Sigma_l \frac{p_l^{(u)}}{(d(l, \ell) + 1)^{\alpha_M}} \end{aligned}$$

DMM_Social: It models user's and user's friends mobility patterns using Pareto distribution as above and the resulting

[3] https://dev.twitter.com/streaming/public

[4] https://dev.twitter.com/rest/public

[5] https://developer.foursquare.com/docs/

[6] https://goo.gl/ayzehx

Dataset	#users	#check-ins	#unique locations	#unique categories	#friendship-links	avg. degree	#users (training records > 10)	Duration	$A(\ell, u)$	cut-off timestamp
FSq'16	119,756	9,317,276	183,225	734	1,308,337	21.85	78,312	Jan'15 - Feb'16	55,884	1/12/2015
FSq'11	11,326	1,385,223	187,218	638	47,164	8.33	11,324	Jan'11 - Dec'11	15,951	1/10/2011
FSq'10	18,107	2,073,740	43,064	624	115,574	12.76	17,369	Mar'10 - Jan'11	4,056	1/12/2010
Brightkite	58,228	4,491,143	772,966	683	214,078	7.35	23,356	Apr'08 - Oct'10	2,642	1/5/2010
Gowalla	196,591	6,442,890	1,280,970	680	950,327	9.66	72,925	Feb'09 - Oct'10	88,865	1/6/2010

Table 3: Statistical properties of the dataset

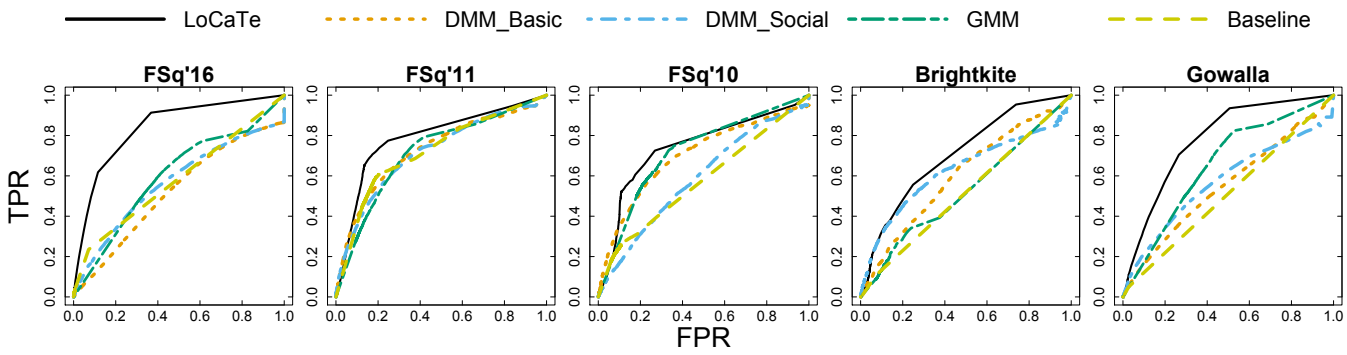


Figure 5: ROC for different influence quantification models

dataset	FSq'16	FSq'11	FSq'10	BrightKite	Gowalla
α	0.90	0.95	0.92	0.93	0.94
β_ν	0.78	0.86	0.85	0.91	0.90

Table 4: α and β_ν values

model is the mixture of individual's distance density and social distance density as follows:

$$P_u(\ell) = \Sigma_l p_l^{(u)} \left[\frac{p(M)}{(d(l,\ell)+1)^{\alpha_M}} + \frac{p(S)}{(d(l,\ell)+1)^{\alpha_S}} \right]$$

where, $p(M)$ and $p(S)$ are mixing components and α_M and α_S are the Pareto distribution parameters learned using individual and social data, respectively.

GMM: It models user's mobility patterns using the Gaussian Mixture Model. Each user's check-in records are represented using several states, each state modeled using Gaussian distribution. In our experiments we choose two states: home and work [Cho *et al.*, 2011; Zhu *et al.*, 2015].

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \right\}$$

$$p(x) = \Sigma_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

where, $\pi_1 ... \pi_k$, are the mixture weights of the states, $\mu_1 ... \mu_k$, the mean of each state and $\Sigma_1 ... \Sigma_k$, the variance of each state.

Baseline: In equation (1) in section 4 we plugin Most Frequent Checkins as the location model, Simple Category Distribution as the category model and average time lag based exponential distribution as the temporal model.

$$P_{\ell,u}(\nu|M) = \left(\alpha \frac{I_\ell}{|M_u|} + (1-\alpha) \frac{I_{C_\ell}}{\Sigma_{i=1}^{|M_u|} |C_i|} \right) \times \exp(-\overline{\Delta t_{u,\nu}}),$$

where, I_ℓ is the number of instances when u has checked-in at ℓ, I_{C_ℓ} is the number of instances when u has checked-in at category set C_ℓ, and $\overline{\Delta t_{u,\nu}}$ is the average of time lag between u and ν check-ins in the training data.

ROC and AUC: Figure 5 shows ROC curves and table 5 shows AUC (area under the curve) of different influence quantification models on different datasets. It can be observed that **LoCaTe** model outperforms DMM_Basic and DMM_Social models quite significantly on FSq'16 dataset, where we have a single category with each location. On the other datasets where we have a set of categories with each location, although we observe that **LoCaTe** model outperforms DMM_Basic and DMM_Social models, but we are able to leverage the category information better in the case of a single category with each location. The efficacy of the **LoCaTe** model is not only contributed by additional knowledge we gain from categories, but also the temporal based user-user correlation modelling, and the adaptive KDE based mobility model (Lo). The **Lo** model provides a better fit to the mobility data as for each testing location the distance around it is determined using the k nearest neighbours (from the training data). On the other hand, the distance based mobility model (DMM) is sensitive to short distances and thus assigns a low probability to locations at larger distances.

dataset \ technique	baseline	GMM	DMM_Basic	DMM_Social	LoCaTe
Fsq'16	0.582	0.599	0.521	0.568	**0.839**
Fsq'11	0.721	0.716	0.727	0.716	**0.789**
Fsq'10	0.575	0.718	0.699	0.588	**0.741**
Brightkite	0.517	0.526	0.601	0.627	**0.707**
Gowalla	0.511	0.654	0.551	0.571	**0.781**

Table 5: AUC(area under the curve) of different influence quantification models over different datasets

6 Conclusion

In this paper, we addressed the problem of influence quantification in the location promotion setting, and proposed a model that incorporates not only the traditional user mobility models but also socially induced temporal correlation of users as well as the affinity of users to a location based on semantics of the location. The resulting *LoCaTe* model has demonstrated more than 54% improvements over state-of-the-art methods over a number of real-world LBSN data with a large number of users and spanning more than a year.

References

[Blei *et al.*, 2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3, January 2003.

[Bouros *et al.*, 2014] Panagiotis Bouros, Dimitris Sacharidis, and Nikos Bikakis. Regionally influential users in location-aware social networks. In *SIGSPATIAL*, 2014.

[Bradley, 1997] Andrew P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recogn.*, 30(7), July 1997.

[Chen *et al.*, 2010] Wei Chen, Chi Wang, and Yajun Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *SIGKDD*, 2010.

[Cho *et al.*, 2011] Eunjoon Cho, Seth A. Myers, and Jure Leskovec. Friendship and mobility: User movement in location-based social networks. In *SIGKDD*, 2011.

[Gao *et al.*, 2012] Huiji Gao, Jiliang Tang, and Huan Liu. Exploring social-historical ties on location-based social networks. In *ICWSM*, 2012.

[Goyal *et al.*, 2010] Amit Goyal, Francesco Bonchi, and Laks V.S. Lakshmanan. Learning influence probabilities in social networks. In *WSDM*, 2010.

[Kempe *et al.*, 2003] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *SIGKDD*, 2003.

[Kempe *et al.*, 2005] David Kempe, Jon Kleinberg, and Éva Tardos. Influential nodes in a diffusion model for social networks. In *ICALP*, 2005.

[Li *et al.*, 2014] Guoliang Li, S Chen, J Feng, Kian-lee Tan, and Ws Li. Efficient location-aware influence maximization. In *SIGMOD*, 2014.

[Lichman and Smyth, 2014] Moshe Lichman and Padhraic Smyth. Modeling human location data with mixtures of kernel densities. In *SIGKDD*, 2014.

[Likhyani *et al.*, 2015] Ankita Likhyani, Deepak Padmanabhan, Srikanta Bedathur, and Sameep Mehta. Inferring and exploiting categories for next location prediction. In *WWW*, 2015.

[Newman, 2005] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 2005.

[Pham and Shahabi, 2016] H. Pham and C. Shahabi. Spatial influence - measuring followship in the real world. In *ICDE*, 2016.

[Wu and Yeh, 2013] Hao-Hsiang Wu and Mi-Yen Yeh. Influential nodes in a one-wave diffusion model for location-based social networks. In *PAKDD*, 2013.

[Yang *et al.*, 2016] Dingqi Yang, Daqing Zhang, and Bingqing Qu. Participatory cultural mapping based on collective behavior data in location-based social networks. *ACM TIST*, 7(3), April 2016.

[Ye *et al.*, 2011] Mao Ye, Peifeng Yin, Wang-Chien Lee, and Dik-Lun Lee. Exploiting geographical influence for collaborative point-of-interest recommendation. In *SIGIR*, 2011.

[Zhang *et al.*, 2012] Chao Zhang, Lidan Shou, Ke Chen, Gang Chen, and Yijun Bei. Evaluating geo-social influence in location-based social networks. In *CIKM*, 2012.

[Zhu *et al.*, 2015] Wen-Yuan Zhu, Wen-Chih Peng, Ling-Jyh Chen, Kai Zheng, and Xiaofang Zhou. Modeling user mobility for location promotion in location-based social networks. In *SIGKDD*, 2015.