Discovering Interpretable Geo-Social Communities for User Behavior Prediction

Hongzhi Yin† Zhiting Hu§ Xiaofang Zhou† Hao Wang‡ Kai Zheng† Quoc Viet Hung Nguyen† Shazia Sadiq†
†*The University of Queensland, School of Information Technology and Electrical Engineering*
§*Language Technologies Institute, Carnegie Mellon University*
‡*SKLCS, Institute of Software, Chinese Academy of Sciences*
†db.hongzhi@gmail.com §zhitinghu@cs.cmu.edu
†{zxf,kevinz,q.nguyen,shazia}@itee.uq.edu.au ‡wanghao@iscas.ac.cn

***Abstract*—Social community detection is a growing field of interest in the area of social network applications, and many approaches have been developed, including graph partitioning, latent space model, block model and spectral clustering. Most existing work purely focuses on network structure information which is, however, often sparse, noisy and lack of interpretability. To improve the accuracy and interpretability of community discovery, we propose to infer users' social communities by incorporating their spatiotemporal data and semantic information. Technically, we propose a unified probabilistic generative model, User-Community-Geo-Topic (*UCGT*), to simulate the generative process of communities as a result of network proximities, spatiotemporal co-occurrences and semantic similarity. With a well-designed multi-component model structure and a parallel inference implementation to leverage the power of multicores and clusters, our UCGT model is expressive while remaining efficient and scalable to growing large-scale geo-social networking data. We deploy UCGT to two application scenarios of user behavior predictions: check-in prediction and social interaction prediction. Extensive experiments on two large-scale geo-social networking datasets show that UCGT achieves better performance than existing state-of-the-art comparison methods.**

I. Introduction

As social networks (e.g., Facebook and Twitter) gain prominence, the first obvious question that comes to a researcher's mind in observing these networks is: how to extract meaningful knowledge from these data? In seeking a response, the network structure proves to be of utmost importance. Identifying high-order structures within networks yields insights into their functional organizations, which in turn contributes more knowledge to support many practical applications, including user behavior predictions and online marketing. Thus, discovering community structures from social networks has attracted considerable research interests. Many approaches have been developed for clustering on graph that serves the purpose of community extraction or discovery, including graph partitioning [11], latent space model [24], block model [1], spectral clustering [17], etc.

However, after investigating multiple social media datasets, an important observation is made that in some cases, reliable social structure information is available for users, while in many cases, these data is unavailable, incomplete or unreliable, either due to the privacy issue or because the user declines to share the true information. Therefore, discovering communities simply based on pure network structure becomes problematic in some scenarios. (1) Consider a spammer in a social network who builds social connections with many users, which leads to a problem to all community discovery methods based on topology structure. (2) Aside from the possible bias in the network topology due to unwanted connections, existing methods also suffer from the lack of interpretation. Given a group of users discovered as a community, a natural question is why they form a community? The methods based on the pure network topology fall short in answering such questions.

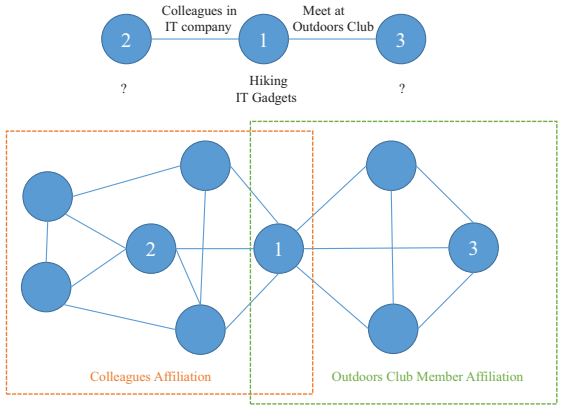


Fig. 1. Toy Example

For the ease of understanding the importance of the interpretability of communities, we provide a toy example in Figure 1. Actor 1 connects Actor 2 because they work in the same IT company, and connects to Actor 3 because they often meet each other in the same outdoors club. Given the label information that Actor 1 is interested in both Hiking and IT Gadgets, can we infer Actor 2 and 3's interests? If we do not know the interpretations of the communities and treat these two connections homogeneously, we guess that both Actors 2 and 3 are also interested in Hiking and IT Gadgets. But if we know how Actor 1 connects to them, it is more reasonable to conjecture that Actor 2 is more interested in IT gadgets and Actor 3 likes Hiking. The accurate inference of users' interests on social networks is critical for social networking advertising, recommendation and search.

To overcome the challenges from low-quality network data and improve the interpretability of discovered communities,

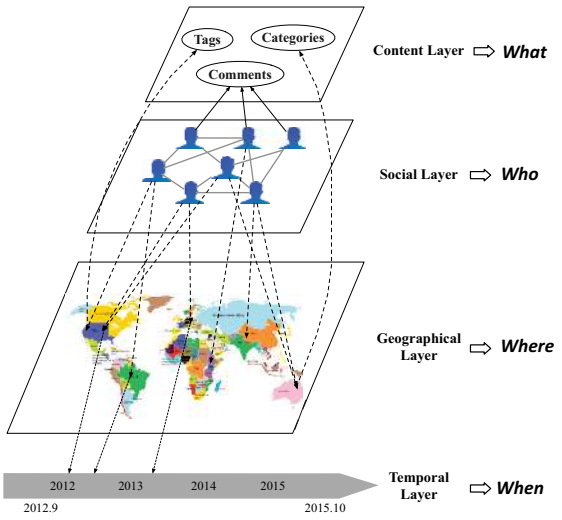


Fig. 2. The Information Layout of Geo-Social Network

we propose to infer users' social communities by incorporating their spatiotemporal data and the associated contents. Recently, the advances in location acquisition and wireless communication technologies enable people to add a location dimension to social networks, fostering a profusion of geo-social networks, such as location-based social networks (LBSNs) and event-based social networks (EBSNs) [16], [22]. LBSNs (e.g., Foursquare, Yelp and Google Place) provide users an online platform to check-in at points of interests (e.g., cinemas, galleries and hotels) and share their life experiences with their online friends. Moreover, newly emerging EBSNs (e.g., Meetup and Plancast) enable users to check-in and share more specific activities/events held in the physical world, ranging from informal get-togethers (e.g., movie nights and dining out) to formal activities (e.g., business meetings). Thus, the geo-social network contains a "4W" (i.e., **who**, **when**, **where** and **what**) information layout, corresponding to four distinct information layers as shown in Figure 2. The dimensions of space-time and semantic imply extensive knowledge about users' behaviors and interests by bridging the gap between online social networks and the physical world, enabling us to better understand the formation of communities.

Specifically, for one thing, users' spatiotemporal data provides a rich source of information for studying users' social communities (e.g., affiliations). Some recent studies [21], [26] showed that there is a correlation between people's spatiotemporal co-occurrences (i.e., two people appear at the same places at the same time) and their social connections, and the co-occurrence frequency (i.e., how often two people co-locate at the same time) has been widely used to measure the strength of social relationship between two people. The intuition is that if two people often co-occur at the same places, there is a good chance that they are socially related and belong to the same community. For another, as suggested in the established social science theory of homophily [19], "birds of a feather flock together". Similarity breeds connections in real-world social networks. In terms of social media, users sharing the same interests and enjoying the same contents tend to belong to the same communities.

In this paper, we approach the problem of community detection using a probabilistic generative model based on Bayesian network, named *User-Community-Geo-Topic* (**UCGT**), which models the formation of communities as a result of network proximity, spatiotemporal co-occurrences and semantic similarity among social actors. To effectively infer the UCGT model, we propose an entropy filtering-based Gibbs sampling method. To demonstrate the potential of our model in practical applications, we evaluate its performance on challenging user behavior predictions: check-in prediction and social interaction prediction. In these two prediction problems, we take advantage of the community members' collective behavior patterns and network structures to overcome the sparsity and volatility of individuals' behaviors and connections.

To summarize, we make the following contributions:

1. **Novelty Perspective**. We identify the problem of community discovery beyond network structure by incorporating both spatiotemporal co-occurrences and semantic similarity. It brings up new insights into the community formation process.
2. **Comprehensive Model**. We propose a Bayesian model UCGT to simulate the generative process of communities as a result of network proximities, spatiotemporal co-occurrences and semantic similarity. To improve the performance of UCGT, we incorporate the idea of entropy filtering to Gibbs sampling.
3. **Scalable Inference**. To adapt to large-scale geo-social data, we develop a scalable parallel implementation of the UCGT by harnessing the powers of multicores and clusters.
4. **Inspiring Prediction and Exploration**. We deploy UCGT to user behavior predictions (i.e., check-in prediction and social interaction prediction) without any modification to the model itself, showing its superiority and significant improvement over existing state-of-the-art methods.

The remainder of the paper is organized as follows. Section 2 formulates the problem and introduces the model. Section 3 describes the inference algorithm and parallel implementation. Section 4 illustrates the applications of UCGT. We report the experimental results in Section 5. We review related literatures in Section 6 and conclude the paper in Section 7.

II. Community Discovery from Geo-Social Network

A. Problem Formulation

Notation. Through this paper, all vectors are column vectors and are denoted by bold lower case letters (e.g., $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$). We use calligraphic letters to represent sets (e.g., $\mathcal{U}$ and $\mathcal{V}$). For simplicity, we use their corresponding normal letters to denote their cardinalities (e.g., $V = |\mathcal{V}|$).

Definition 1: **(Interaction Network).** As a mathematical abstraction, we define the interaction network as a directed graph $\mathcal{G} = (\mathcal{U}, \mathcal{E})$, where $\mathcal{U}$ is a set of nodes/users, and $\mathcal{E}$ is a set of edges. The link set $\mathcal{E}$ denotes interactions between users and can be derived from various types of user interactions such as following, retweeting, emailing and co-authoring. A directed link $(u,\ u') \in \mathcal{E}$ indicates that there

TABLE I
NOTATIONS OF PARAMETERS

Variable	Interpretation
$\boldsymbol{\pi}_u$	the community memberships of user u, expressed by a multinomial distribution over communities
$\boldsymbol{\theta}_c$	the interests of community c, expressed by a multinomial distribution over topics
$\boldsymbol{\vartheta}_c$	a multinomial distribution over spatial items specific to community c
$\boldsymbol{\phi}_z$	a multinomial distribution over words specific to topic z
$\psi_{c,c'}$	the general interaction strength between communities c and c'
λ_0, λ_1	Beta priors on $\boldsymbol{\psi}$
$\boldsymbol{h}_{c,v}$	the bandwidth vector of the kernel function specific to (c, v)
$\alpha, \beta, \gamma, \eta$	Dirichlet priors to multinomial distributions $\boldsymbol{\theta_c}$, $\boldsymbol{\phi_z}$, $\boldsymbol{\pi_u}$ and $\boldsymbol{\vartheta_c}$, respectively

exists communication from user u to u'. We use $\mathcal{E}_u$ to denote the set of links from u to other users.

Definition 2: **(Spatial Item)** A spatial item $v \in \mathcal{V}$ is defined as a uniquely identified specific site (e.g., a restaurant or a cinema) or an event (e.g., a conference or an exhibition).

Definition 3: **(Check-in Activity)** A check-in activity is made of a four-tuple $(u,\ v,\ t,\ \mathcal{D})$ that means user u checks-in at spatial item v at time t. $\mathcal{D}$ denotes a collection of words extracted from user comments or the descriptions/tags associated with v. For each user u, we use $\mathcal{L}_u$ to denote the collection of her check-in activities.

A community is a collection of users with more intense interactions amongst its members than other users. It can be characterized not only by social link structures, but also geographical and topical preferences. Therefore, we associate each community with a distribution over topics representing its interests, a distribution over spatial items indicating its spatial activities, and an interaction vector representing its interaction strengths with other communities. A formal definition of community is given below.

Definition 4: **(Community).** A community $c \in \mathcal{C}$ has three components: (1) a multinomial distribution over topics $\boldsymbol{\theta}_c$, where each component $\theta_{c,z}$ represents the probability that community c is interested in topic z; (2) a multinomial distribution over spatial items $\boldsymbol{\vartheta}_c$, where each component $\vartheta_{c,v}$ represents the probability that community c visits spatial item v; (3) a probability vector $\boldsymbol{\psi}_c$, where each component $\psi_{c,c'}$ is the mean of a Bernoulli distribution representing the interaction probability between communities c and c'.

In social networks, users usually have multiple roles and belong to multiple affiliations [28]. We therefore employ the *mixed-membership* approach: each user u is associated with a multinomial distribution over communities $\boldsymbol{\pi}_u$, where $\pi_{u,c}$ indicates her affiliation degree to community c.

Definition 5: **(Topic)**. Given a collection of words $\mathcal{W}$, a topic z is defined as a multinomial distribution over $\mathcal{W}$, i.e., $\boldsymbol{\phi}_z = \{\phi_{z,w} : w \in \mathcal{W}\}$ where each component $\phi_{z,w}$ denotes the probability of word w generated from topic z. Generally, a topic is a semantic-coherent soft cluster of words.

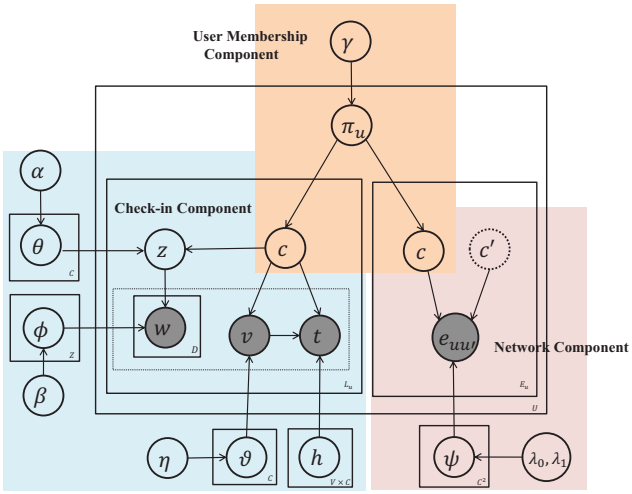


Fig. 3. Graphical Representation of UCGT. The latent variable c' is represented by a dashed circle since it is drawn from $\pi_{u'}$ which is not shown in the graph model.

B. Model Structure

To detect user communities from geo-social networking data, we propose a probabilistic generative model, User-Community-Geo-Topic (UCGT), which jointly models both users' check-in behaviors $\mathcal{L}_u$ and interaction behaviors $\mathcal{E}_u$. It considers the formation of communities as a result of spatiotemporal co-occurrences, semantic similarity and network proximity among social actors.

Figure 3 shows the graphical structure of UCGT, and the notations of the model are listed in Table I. By jointly considering the two types of user behaviors with properly separated generative process, UCGT naturally combines check-in data and network data while still keeping the model tractable. Specifically, there are three components in UCGT: the *check-in component* captures the spatiotemporal patterns of communities and uncovers the semantic topics; the *network component* accounts for the link structure; and the *user membership component* models user membership to communities, which also serves to seamlessly unify the other two components.

User Membership Component. Generally, users have multiple affiliations in the real world. Correspondingly, users in a social network also have multiple community memberships. We associate each user u with a community probability vector $\boldsymbol{\pi}_u$. A community c is assigned to a check-in activity $(u,\ v,\ t,\ \mathcal{D}) \in \mathcal{L}_u$, denoting the community membership (e.g., the role) of user u when visiting spatial item v. In addition, a community pair (c, c') is assigned to a positive link $e_{u,u'} \in \mathcal{E}_u$, which denotes the community memberships of users u and u' when they build the social link.

Check-in Component. Since each check-in activity $(u,\ v,\ t,\ \mathcal{D}) \in \mathcal{L}_u$ contains both spatiotemporal and semantic information, this component consists of two subcomponents: *Spatiotemporal Subcomponent* and *Semantic Subcomponent*.

Spatiotemporal Subcomponent. Users with the same roles or affiliations tend to visit the same places and attend the same events. To model this observation, each community c

is associated with a probabilistic distribution $\boldsymbol{\vartheta}_c$ over spatial items, and each component $\vartheta_{c,v}$ denotes the probability of community c visiting spatial item v. In addition, the activity time of a community's members tend to be close to each other. Thus, for simplicity and speed, we propose to use Kernel Density Estimation (KDE) method to model the continuous time density of users from the same community c visiting the same spatial item v.

KDE is a non-parametric model for estimating density from sample points. Following the kernel density model, we define the probability of a user from community c visiting spatial item v at time t as in Equation 1.

$$P(t|\boldsymbol{h}_{c,v}) = \frac{1}{|\mathcal{T}_{c,v}|} \sum_{t' \in \mathcal{T}_{c,v}} K_{h^{t'}_{c,v}}(t - t') \tag{1}$$

where $\mathcal{T}_{c,v}$ is a collection of time stamps $\mathcal{T}_{c,v}$ at which all other users from community c visit POI v, and $K_{h^{t'}_{c,v}}(t-t')$ is defined as follows:

$$K_{h^{t'}_{c,v}}(t - t') = \frac{1}{2\pi h^{t'}_{c,v}} \exp\left(\frac{-(t-t')^2}{2{h^{t'}_{c,v}}^2}\right) \tag{2}$$

where $K(\cdot)$ is a Gaussian kernel function, and $h^{t'}_{c,v} > 0$ is an adaptive bandwidth parameter [15] for the data point t'. In our model, we choose to use an adaptive bandwidth parameter $h^{t'}_{c,v}$ for each data point t', instead of a fixed bandwidth parameter $h_{c,v}$ for all data points in $\mathcal{T}_{c,v}$, because it is difficult to choose a suitable common bandwidth parameter $h_{c,v}$ for all data points in $\mathcal{T}_{c,v}$. $P(t|h_{c,v})$ is highly sensitive to the value of the bandwidth $h_{c,v}$, producing densities that are sharply peaked around the data points in $\mathcal{T}_{c,v}$ when $h_{c,v}$ is too small, and leading to an overly smooth estimate that may omit important structure in the data (such as multiple modes) when $h_{c,v}$ is too large.

Semantic Subcomponent. Users with the same communities tend to have the same interests and enjoy the same contents. In each check-in activity $(u,\ v,\ t,\ \mathcal{D})$, $\mathcal{D}$ is a bag of words describing the check-in contents. We therefore associate $\mathcal{D}$ with a latent variable z generated from the community's interest distribution $\boldsymbol{\theta}_c$ to indicate its topic. Note that in the traditional topic models [2] such as LDA, a document contains a mixture of topics, and each word has a hidden topic label. This is reasonable for long documents. However, the document $\mathcal{D}$ in a check-in activity (e.g., tags of a restaurant) is usually short, and is most likely to be about a single topic. Thus in UCGT, all the words in $\mathcal{D}$ are assigned with a single topic z, and they are generated from the same word distribution $\boldsymbol{\phi}_z$.

Network Component. We use pairwise community Bernoulli distributions $\boldsymbol{\psi}$ to model the presence and absence of links between pairs of users. For each link (u, u'), a boolean indicator $e(u, u')$ is drawn from $\psi_{c,c'}$ which represents the interaction strength between community c and c'.

Social networks are typically sparse, thus we only model positive links: the variables c and c' exist if and only if $e(u, u') \in \mathcal{E}_u$. As in [9], the negative links $e(u, u')$ are implicitly modeled in a Bayesian fashion: we use a $Beta(\lambda_0, \lambda_1)$ prior on each $\psi_{c,c'}$, and set $\lambda_0 = \kappa \cdot ln(n_{neg}/C^2)$ and $\lambda_1 = 0.1$, where $n_{neg} = U(U-1) - \sum_u E_u$ is the number of negative links, and κ is a tunable weight. In this way, we reduce large amount of computation and achieve linear complexity on network modeling, as shown in Section III-A.

Algorithm 1: Probabilistic generative process in UCGT

```
for each community c ∈ C do
    Sample the distribution over topics θ_c ~ Dirichlet(·|α);
    Sample the distribution over spatial items
    ϑ_c ~ Dirichlet(·|η);
    for each community c' ∈ C do
        Sample community-community link probability
        ψ_{c,c'} ~ Beta(λ_0, λ_1);
    end
end
for each topic z ∈ Z do
    Sample the distributions over words φ_z ~ Dirichlet(·|β);
end
for each user u ∈ U do
    Sample the distribution over communities
    π_u ~ Dirichlet(·|γ);
    for each check-in activity (u, v, t, D) ∈ L_u do
        Sample a community indicator c ~ Multi(π_u);
        Sample a topic indicator z ~ Multi(θ_c);
        Sample a spatial item v ~ Multi(ϑ_c);
        Sample time t according to Equation (1);
        for each word w ∈ D do
            Sample word w ~ Multi(φ_z);
        end
    end
    for each positive link e(u, u') ∈ E_u do
        Sample a community indicator c ~ Multi(π_u);
        Sample a community indicator c' ~ Multi(π_u');
        Sample social link e_{u,u'} ~ Bernoulli(ψ_{c,c'});
    end
end
```

Joint Modeling of Three Components. Note that, to avoid overfitting, we place a Dirichlet prior over each multinomial distribution. For example, Dirichlet prior parameter α is incorporated for $\boldsymbol{\theta}_c$, as follows:

$$P(\boldsymbol{\theta}_c|\alpha) = \frac{\Gamma(\sum_z \alpha)}{\prod_z \Gamma(\alpha)} \prod_z \theta_{c,z}^{\alpha-1}, \tag{3}$$

where $\Gamma(\cdot)$ is the gamma function. Similarly, priors over $\boldsymbol{\vartheta}_c$, $\boldsymbol{\phi}_z$ and $\boldsymbol{\pi}_u$ are imposed with parameters η, β and γ, respectively. Based on the three components described above, we obtain the joint distribution of the observed and hidden variables as described in Equation 4, where we use different colors to represent different components (User Membership Component, Check-in Component, and Network Component).

C. Generative Process

The generative process is summarized in Algorithm 1. Consider a user u who visits spatial items and interacts with others. When she visits a spatial item v, she first selects the community membership c (e.g., her role) by her community distribution $\boldsymbol{\pi}_u$, then selects a topic z by the community's topic distribution $\boldsymbol{\theta}_c$. With the chosen community c, spatial item v is generated from the community's spatial distribution $\boldsymbol{\vartheta}_c$. With the chosen community c and spatial item v, time t is generated from the community's temporal distribution w.r.t. spatial item v. With the chosen topic z, words in $\mathcal{D}$ are generated from the topic's word distribution. On the other hand, when she interacts with another user u', a community is

$$
\begin{aligned}
&P(\boldsymbol{v}, \boldsymbol{t}, \mathcal{D}, \boldsymbol{e}, \boldsymbol{c}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\vartheta}, \boldsymbol{\psi} | \alpha, \beta, \gamma, \eta, \lambda_0, \lambda_1, \boldsymbol{h}) \\
=&P(\boldsymbol{\pi}|\gamma)P(\boldsymbol{c}|\boldsymbol{\pi})P(\boldsymbol{\vartheta}|\eta)P(\boldsymbol{v}|\boldsymbol{c}, \boldsymbol{\vartheta})P(\boldsymbol{t}|\boldsymbol{c}, \boldsymbol{v}, \boldsymbol{h})P(\boldsymbol{\theta}|\alpha)P(\boldsymbol{z}|\boldsymbol{\theta})P(\boldsymbol{\phi}|\beta)P(\mathcal{D}|\boldsymbol{z}, \boldsymbol{\phi})P(\boldsymbol{\psi}|\lambda_0, \lambda_1)P(\boldsymbol{e}|\boldsymbol{c}, \boldsymbol{\psi})
\end{aligned} \tag{4}
$$

sampled for each of them according to their own community distributions, and the link is formed by the community-community interaction strength $\psi_{c,c'}$.

III. Inference & Implementation

In this section, we first present the basic inference algorithm using Gibbs sampling method. To improve the performance of Gibbs sampling, we incorporate the idea of entropy filtering. Then, to adapt to the large-scale geo-social data, we develop a parallel implementation of UCGT to ensure high scalability.

A. Gibbs Sampling

Exact inference of UCGT model is difficult due to the intractable normalizing constant of the posterior distribution. We therefore adopt collapsed Gibbs sampling [33], [10] for approximate inference. As a widely used Markov chain Monte Carlo (MCMC) algorithm, Gibbs sampling iteratively samples latent variables (i.e., $\{\boldsymbol{c}, \boldsymbol{z}\}$ in UCGT) from a Markov chain, whose stationary distribution is the posterior. The samples can therefore be used to estimate the distributions of interest (i.e., $\{\boldsymbol{\theta}, \boldsymbol{\vartheta}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\psi}\}$). As for the hyperparameters α, β, γ and η, for simplicity, we take a fix value, i.e., $\alpha = 50/Z$, $\gamma = 50/C$ and $\beta = \eta = 0.01$, following the studies [33], [10], where Z and C are the numbers of topics and communities, respectively.

At each iteration of our Gibbs sampler, for each check-in activity, we sample both the corresponding community indicator c and the topic indicator z; for each link $e_{u,u'}$, we sample the corresponding community indicators c and c'. Due to space constraints, we show only the derived Gibbs sampling formulas, omitting the detailed derivation process.

For each user check-in activity $(u,\ v,\ t,\ \mathcal{D})$, we first sample community c according to the following posterior probability:

$$
\begin{aligned}
&P(c|\boldsymbol{c}_{\neg}, \boldsymbol{z}, \boldsymbol{v}, \boldsymbol{t}, u, \cdot) \propto (n^{\neg}_{u,c} + \gamma) \\
&\times \frac{n^{\neg}_{c,z} + \alpha}{\sum_{z'}(n^{\neg}_{c,z'} + \alpha)} \frac{n^{\neg}_{c,v} + \eta}{\sum_{v'}(n^{\neg}_{c,v'} + \eta)} P(t|\hat{\boldsymbol{h}}_{c,v})
\end{aligned} \tag{5}
$$

where $\boldsymbol{c}_{\neg}$ represents community assignments for all check-in records except the current one; $n_{u,c}$ is the number of times that latent community c is sampled from user u; $n_{c,z}$ is the number of times that topic z is generated by community c; $n_{c,v}$ is the number of times that spatial item v is generated by community c; and the number $n^{\neg}$ with superscript $\neg$ denotes a quantity excluding the current instance. For the adaptive bandwidth parameter $\boldsymbol{h}_{c,v}$, we let $h^{t'}_{c,v}$ be the Euclidean distance of t' to its k-th nearest neighbor in $\mathcal{T}_{c,v}$, following recent study [15]. $\mathcal{T}_{c,v}$ is the collection of time stamps at which users from community c visit spatial item v, except the current one.

After c is sampled, we sample topic z conditioned on the newly sampled c, as follows:

$$
P(z|\boldsymbol{z}_{\neg}, \boldsymbol{c}, \boldsymbol{d}, \cdot) \propto (n^{\neg}_{c,z} + \alpha) \prod_{w \in \mathcal{D}} \frac{n^{\neg}_{z,w} + \beta}{\sum_{w'}(n^{\neg}_{z,w'} + \beta)} \tag{6}
$$

where $n_{z,w}$ is the number of times that word w is generated from topic z.

For each social link $e(u, u')$, we sample communities c and c' according to:

$$
\begin{aligned}
&P(c, c'|e_{u,u'} = 1, \boldsymbol{e}_{\neg}, \boldsymbol{c}, \cdot) \\
&\propto (n^{\neg}_{u,c} + \gamma)(n^{\neg}_{u',c'} + \gamma) \frac{n^{\neg}_{c,c'} + \lambda_1}{n^{\neg}_{c,c'} + \lambda_1 + \lambda_0}
\end{aligned} \tag{7}
$$

where $n^{\neg}_{c,c'}$ is the number of positive links, with current link excluded, whose community indicators are c and c'.

Inference Framework. After a sufficient number of sampling iterations, the approximated posteriors can be used to estimate parameters by examining the counts of $\boldsymbol{z}$ and $\boldsymbol{c}$ assignments to check-in records and social links. The detailed inference framework is shown in Algorithm 2 where $\mathcal{L} = \bigcup_{u \in \mathcal{U}} \mathcal{L}_u$. We first randomly initialize the topic and community assignments for each check-in record (Lines 2-4) and community-community assignments for each positive link (Lines 5-7). Afterwards, in each iteration, sampling Formulas (5, 6) are utilized to update the community and topic assignments for each check-in record $(u,\ v,\ t,\ \mathcal{D})$ (Lines 10-13), and Formula (7) is used to update the community pair assignment for each positive link $e(u, u')$ (Lines 14-16). For speed and efficiency, we use the widely adopted late update strategy [25] to defer the update of adaptive bandwidth parameters to the end of each iteration (Lines 17-21). The iteration is repeated until convergence (Lines 9-26). In addition, a burn-in process is introduced in the first several hundreds of iterations to remove unreliable sampling results (Lines 22-25). We also introduce the sample lag (i.e., the interval between samples after burn-in) to sample only periodically thereafter to avoid correlations between samples.

Time Complexity. We now analyze the time complexity of our inference algorithm. It is shown that the devised algorithm scales linearly in terms of the size of data, i.e., the number of check-in records and positive links. In each iteration, the communities of user check-in activities are first sampled. Since all the counters (e.g., $n_{c,z}$ and $n_{c,v}$) and adaptive bandwidth parameters involved in Equation (5) can be cached and updated in constant time for each community c being sampled, Equation (5) can be calculated in constant time. Thus, sampling all c takes linear time w.r.t. the number of check-in records. Similarly, sampling all z by Equation (6) is also linear to the number of check-in records. Next, we sample community indicators c and c' using Equation (7). Since we have implicitly modeled negative links in Bayesian prior (i.e., the Beta prior for $\psi_{c,c'}$), we only need to sample c and c' for positive links $e(u, u')$. Hence the complexity is reduced from quadratic (w.r.t. the number of users) to linear (w.r.t. the number of positive links). It significantly saves computation cost due to the sparseness of networks.

B. Gibbs Sampling With Entropy Filtering

Conventionally, the social relationship has been measured by meeting frequency (i.e., how often two people co-locate

Algorithm 2: Inference Framework of UCGT

Input: user check-in collection $\mathcal{L}$, social link collection $\mathcal{E}$, number of iteration I, number of burnin I_b, sample lag I_s, Priors α, γ, β, η, λ_0 and λ_1
Output: estimated parameters $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\vartheta}}$, $\hat{\boldsymbol{\phi}}$, $\hat{\boldsymbol{\pi}}$, $\hat{\boldsymbol{\psi}}$, $\hat{\boldsymbol{h}}$

Create temporary variables $\boldsymbol{\theta}^{sum}$, $\boldsymbol{\vartheta}^{sum}$, $\boldsymbol{\phi}^{sum}$, $\boldsymbol{\pi}^{sum}$ and $\boldsymbol{\psi}^{sum}$, and initialize them with zero;
for *each check-in activity* $(u,\ v,\ t,\ \mathcal{D}) \in \mathcal{L}$ **do**
  Sample community and topic randomly;
end
for *each positive link* $e(u,\ u') \in \mathcal{E}$ **do**
  Sample community pair randomly;
end
Initialize variable *count* with zero;
for $iteration = 1$ *to* I **do**
  for *each check-in activity* $(u,\ v,\ t,\ \mathcal{D}) \in \mathcal{L}$ **do**
    Sample community c according to Equation (5);
    Sample topic z according to Equation (6);
  end
  for *each positive link* $e(u,\ u') \in \mathcal{E}$ **do**
    Sample community pair (c, c') according to Equation (7);
  end
  for *each community* $c \in \mathcal{C}$ **do**
    for *each spatial item* v *associated with* c **do**
      Update the adaptive bandwidth parameters $\hat{\boldsymbol{h}}_{c,v}$;
    end
  end
  if $(iteration > I_b)$ *and* $(iteration\ mod\ I_s == 0)$ **then**
    $count = count + 1$;
    Update $\boldsymbol{\theta}^{sum}$, $\boldsymbol{\vartheta}^{sum}$, $\boldsymbol{\pi}^{sum}$ $\boldsymbol{\phi}^{sum}$ and $\boldsymbol{\psi}^{sum}$ as follows:

$$\theta^{sum}_{c,z} += \frac{n_{c,z} + \alpha}{\sum_{z'}(n_{c,z'} + \alpha)}$$

$$\vartheta^{sum}_{c,v} += \frac{n_{c,v} + \eta}{\sum_{v'}(n_{c,v'} + \eta)}$$

$$\pi^{sum}_{u,c} += \frac{n_{u,c} + \gamma}{\sum_{c'}(n_{u,c'} + \gamma)}$$

$$\phi^{sum}_{z,w} += \frac{n_{z,w} + \beta}{\sum_{w'}(n_{z,w'} + \beta)}$$

$$\psi^{sum}_{c,c'} += \frac{n_{c,c'} + \lambda_1}{n_{c,c'} + \lambda_1 + \lambda_0}$$

  end
end
Return model parameters $\hat{\boldsymbol{\theta}} = \frac{\boldsymbol{\theta}^{sum}}{count}$, $\hat{\boldsymbol{\vartheta}} = \frac{\boldsymbol{\vartheta}^{sum}}{count}$, $\hat{\boldsymbol{\pi}} = \frac{\boldsymbol{\pi}^{sum}}{count}$, $\hat{\boldsymbol{\phi}} = \frac{\boldsymbol{\phi}^{sum}}{count}$, $\hat{\boldsymbol{\psi}} = \frac{\boldsymbol{\psi}^{sum}}{count}$ and $\hat{\boldsymbol{h}}$;

at the same time) [7]. However, we argue that these meeting events should not be treated equally, and propose to consider the popularity information of spatial items. Some spatial items are very popular and frequently visited by many people, such as the downtown in the city and a popular restaurant, whereas other locations are more specific only to a few people, such as a private house. In a popular public place, it is more likely for two strangers to co-locate by coincidence. Thus, such meeting events are less indicative for a relationship. In contrast, a meeting event in a private place often indicates a strong social relationship. Therefore, the Gibbs sampling may yield poor performance by including many popular spatial items.

In light of this, we incorporate the idea of entropy filtering into the Gibbs sampling, resulting in Entropy Filtering-Gibbs sampling algorithm (EnF-Gibbs) which can automatically remove check-in records associated with the non-informative spatial items based on entropy measure [21]. During the procedure of EnF-Gibbs sampling, the algorithm keeps and maintains a set of spatial items called *TrashCan* that are not informative. Intuitively, the spatial items that are popular among many communities are put into *TrashCan*. After I_b times of iterations in Algorithm 2, we start to ignore the check-in records that contain spatial items that are already in the *TrashCan*. We quantify the informativeness of a spatial item v by its entropy defined in Equation 8. If the entropy of a spatial item v is larger than a threshold and not yet in *TrashCan*, we add it into *TrashCan*.

$$Entropy(v) = -\sum_{c \in \mathcal{C}} P(v|c) log P(v|c) \tag{8}$$

where $P(v|c)$ is computed as follows:

$$P(v|c) = \frac{n_{c,v} + \eta}{\sum_{v'} n_{c,v'} + \eta} \tag{9}$$

In the above equation, the count $n_{c,v}$ is dynamically updated in the Gibbs sampling procedure, thus, *TrashCan* is also automatically updated.

C. Parallel Implementation

To apply UCGT to large-scale geo-social data, parallelizing the inference algorithm of UCGT is inevitable. Each user's check-in collection $\mathcal{L}_u$ or social link collection $\mathcal{E}_u$ can be treated as a document. Many recently developed parallel computation frameworks for machine learning, such as MapReduce [20], make a parallel implementation of Gibbs sampling as follows. Given M processors, a collection of documents are partitioned into M blocks which are processed independently. After each pass the document statistics are synchronized in a separate step. However, on multiprocessor system this approach automatically leads to an $\mathcal{O}(M)$ increase in allocated memory and thereby out-of-memory situations when many cores are involved.

To ensure the scalability of our model in terms of memory and computation time, we first implemented a parallel UCGT inference algorithm by leveraging the strength of *multicore* processors. The state of the sampler comprises the community-topic count matrix $n_{c,z}$, community-spatial item count matrix $n_{c,v}$, topic-word count matrix $n_{z,w}$, community-community count matrix $n_{c,c'}$ and user-community count matrix $n_{u,c}$. The key idea for parallelizing the sampler in the multicore setting is that the first four count matrices (which we will refer to as *state* of the system) change only little given the changes in a single check-in collection $\mathcal{L}_u$ or a single social link collection $\mathcal{E}_u$. Hence, we can assume that $n_{c,z}$, $n_{c,v}$, $n_{z,w}$ and $n_{c,c'}$ are essentially constant while sampling communities and topics for a single $\mathcal{L}_u$ or $\mathcal{E}_u$. This means that there is no need to update $n_{c,z}$, $n_{c,v}$, $n_{z,w}$ and $n_{c,c'}$ during the sampling process and we can defer this action to a separate synchronization thread which takes actions once a single $\mathcal{L}_u$ or $\mathcal{E}_u$ has been entirely processed. Consequently, we can execute a large number of sampling threads simultaneously to process a single $\mathcal{L}_u$ or $\mathcal{E}_u$, which is called intra-document parallelization. Thus, we only need a single set of state variables (e.g., $n_{c,z}$, $n_{c,v}$, $n_{z,w}$ and $n_{c,c'}$) per computer rather than per core. This

dramatically reduces the memory requirements per machine compared with traditional inter-document parallelization.

To further speed up the model training and adapt to large-scale geo-social data, we use a distributed implementation based on the blackboard architecture in [25] to take advantage of the power of *clusters*. The key idea is to have a global consensus of the state variables and to reconcile their values *one entry in the count matrix at a time* asynchronously for all samplers. The advantage is that no synchronization is required between samplers/nodes. When processing $\mathcal{L}_u$, community-topic, community-spatial item and topic-word count matrices are shared across users and maintained in a distributed hash table using `memcached`[1]. The user-community count is user-specific and can be maintained locally in each node. We distribute all users' check-in data across M nodes. We apply the intra-document parallelization strategy for each node to execute multiple foreground threads to sample communities and topics according to Equations 5 and 6. Besides, each node executes a background thread that synchronizes its local copies of community-topic, community-spatial item and topic-word count matrices with the global copy in memcached. When processing user network data $\mathcal{E}_u$, user-community and community-community count matrices also need to be maintained in the memcached for sharing.

IV. User Behavior Prediction

In this section, we deploy UCGT to two types of user behavior predictions: check-in and social interaction predictions. Our solution to user behavior predictions takes advantage of the community members' collective behavior patterns which are stable and predictable. In contrast, traditional methods such as matrix factorization and collaborative filtering can be ineffective due to the volatility of individual's behaviors and the sparsity of individual's check-ins and social connections.

A. Check-in Prediction

Based on the learnt model parameters $\mathbf{\Psi} = \{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\vartheta}}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\pi}}, \hat{\boldsymbol{h}}\}$ in the UCGT, given a target user u and time t, we estimate the probability of user u to visit each unvisited spatial item v, as follows:

$$\begin{aligned} P(v|u,t,\hat{\mathbf{\Psi}}) &= \frac{P(v,t|u,\hat{\mathbf{\Psi}})}{\sum_{v'} P(v',t|u,\hat{\mathbf{\Psi}})} \\ &\propto P(v,t|u,\hat{\mathbf{\Psi}}) \\ &= \sum_c \hat{\pi}_{u,c}\hat{\vartheta}_{c,v} P(t|\hat{\boldsymbol{h}}_{c,v}) \sum_z \hat{\theta}_{c,z} \Big(\prod_{w\in\mathcal{D}} \hat{\phi}_{z,w} \Big)^{\frac{1}{D}} \end{aligned} \tag{10}$$

where $\mathcal{D}$ is a collection of words extracted from v's descriptions/tags or comments from other users. We adopt geometric mean for the probability of topic z generating word set $\mathcal{D}$, i.e., $P(\mathcal{D}|z,\hat{\boldsymbol{\phi}}) = \Big(\prod_{w\in\mathcal{D}} \hat{\phi}_{z,w} \Big)^{\frac{1}{D}}$, considering that the number of words for different spatial items may be different. After computing the check-in probability for each unvisited item v, we can select the top-N ones with the highest probabilities as recommendations.

[1]http://memcached.org/

TABLE II
Basic statistics of Foursquare and Douban Event datasets

	Foursquare	Douban Event
# of users	4,163	295,395
# of spatial item	21,142	350,629
# of check-ins	483,813	18,928,476
# of social links	32,512	30,068,754
time span	Dec 2009-Jul 2013	Sep 2005-Dec 2012

B. Social Interaction Prediction

Social interaction prediction (i.e., link prediction) is defined to estimate the probability of a link. The probability of a link from user u to u' is computed as follows:

$$\begin{aligned} P(u'|u,\hat{\mathbf{\Psi}}) &= \sum_c \sum_{c'} P(c|u)P(c'|c)P(u'|c') \\ &\propto \sum_c \sum_{c'} P(c|u)P(c'|c)P(c'|u')P(u') \\ &\propto \sum_c \sum_{c'} \hat{\pi}_{u,c}\hat{\psi}_{c,c'}\hat{\pi}_{u',c'}P(u') \end{aligned} \tag{11}$$

where $P(u')$ denotes the prior probability of user u', indicating her activeness. In our experiment, we use the normalized number of her followers to represent $P(u')$.

V. Experiments

In this section, we evaluate the performance of the proposed UCGT model in terms of prediction accuracy and model training efficiency on two real-world large-scale datasets. We also use a case study to qualitatively demonstrate its effectiveness in detecting communities.

A. Datasets

Our experiments are conducted on two real geo-social networking datasets: Foursquare and Douban Event. Their basic statistics are shown in Table II.

Foursquare. Foursquare is a popular location-based social network. The dataset used in our experiment contains the check-in history of 4,163 users who live in the California, USA. For each user, it contains her social networks, check-in POI IDs, location of each check-in POI in terms of latitude and longitude, check-in time and the contents of each check-in POI. Each check-in is stored as *user-ID, POI-ID, check-in time, check-in content*, and each record in the social network is stored as *user-ID, friend-ID*.

Douban Event. Douban Event[2] is a Chinese online social event service that helps people publish and participate in social events which are held offline. On Douban Event, a social event is created by a user or an organizer by specifying when, where and what the event is. Then, other users express their intent to join the event by online check-in. We collected a real dataset for events and users by crawling Douban Event from Sep 2005 to Dec 2012. For each event, its content introduction, geographical location, start time information, and a list of registered users for attending were collected. For each user, we acquired her event attendance list and social friend list.

[2]http://www.douban.com/events

B. Performance in User Check-in Prediction

We compare our proposed UCGT models (UCGT-Gibbs in Section III-A and UCGT-EnF-Gibbs in Section III-B) to some representative user check-in (or mobility) prediction methods in the geo-social networks. Table III lists the characters of these methods where we use UCGT-G and UCGT-EG to denote UCGT-Gibbs and UCGT-EnF-Gibbs, respectively. SVDFeature is a *general* feature-based factorization model, while Geo-SAGE, CBPF and Rank-GeoFM are designed for *spatial item*, *event* and *POI* recommendations, respectively.

SVDFeature. SVDFeature [3] is a feature-based matrix factorization model. We implement it by incorporating more side information beyond the user-item matrix, including item content, item location and check-in time. A user-user interaction matrix is also incorporated, inspired by [18].

Geo-SAGE. Geo-SAGE [27] is a geographical sparse additive generative model for predicting user check-in behaviors. This model considers both user's personal interests and the preferences of the crowd traveling or living in the same target region, by exploiting both the co-occurrence pattern of spatial items and the content of spatial items.

CBPF. CBPF [36] is a collective Bayesian Poisson factorization model for event prediction/recommendation. CBPF takes Bayesian Poisson factorization as its basic unit to model user response to events, social relation, and content text separately. Then, it further jointly connects these units by the idea of standard collective matrix factorization model.

Rank-GeoFM. Rank-GeoFM [13] is a ranking based geographical factorization method for POI recommendation, which processes the users' check-in data as implicit feedback information. This model incorporates both geographical influence and temporal influence.

Evaluation Method. To evaluate the prediction accuracy of our models, we first rank the check-in records in each $\mathcal{L}_u$ according to their check-in timestamps. Then, we use the 80-th percentile as the cut-off point so that check-ins before this point will be used for training and the rest are for testing. We adopt the measurement Accuracy@N proposed in [33]. Specifically, for each check-in $(u,\ v,\ t,\ \mathcal{D})$ in $\mathcal{L}_{test}$: 1) We compute the probability of u visiting v and all other spatial items which are within the circle with center v and radius $100km$ and unvisited by u previously, instead of all available ones, since only those ones which are geographically close to v are comparable with v. This design can effectively simulate the local competition effect and user behavior of choices. 2) We form a ranked list by ordering all of these spatial items according to their checked-in probabilities. Let p denote the position of v within this list. The best result corresponds to the case where v precedes all the unvisited ones (that is, $p = 1$). 3) We form a top-N prediction list by picking the N top ranked ones from the list. If $p \leq N$, we have a hit (i.e., the ground truth v is successfully predicted). Otherwise, we have a miss.

The computation of Accuracy@N proceeds as follows. We define hit@N for a single test case as either the value 1, if the ground truth item v appears in the top-N results, or the value 0, if otherwise. The overall Accuracy@N is defined by averaging over all test cases:

$$Accuracy@N = \frac{\#hit@N}{L_{test}}$$

where $\#hit@N$ denotes the number of hits in the test set, and L_{test} is the number of all test cases in $\mathcal{L}_{test}$.

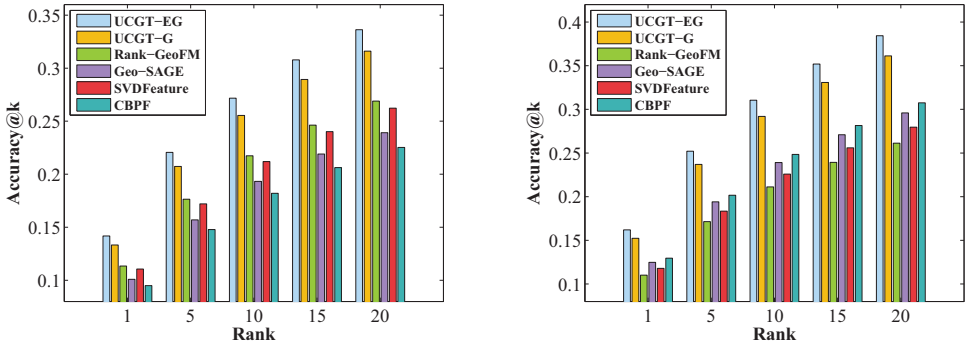
TABLE III
FEATURES OF DIFFERENT METHODS.

Methods \ Features	Spatial	Temporal	Social	Textual
UCGT-G	•	•	•	•
UCGT-EG	•	•	•	•
SVDFeature	•	•	•	•
Geo-SAGE	•			•
CBPF	•		•	•
Rank-GeoFM	•	•		
PMTLM			•	•
COLD		•	•	•
EBM	•	•		

Experimental Results. Figure 4 presents the results of all comparison methods. Clearly, our proposed UCGT models outperform other competitor models significantly and consistently. The main reason is that our UCGT models take advantage of the community members' collective check-in behavior patterns to overcome the sparsity and volatility of individuals' check-in behaviors. Another reason is that UCGT models have the powerful modeling ability to exploit all the information associated with each user in a geo-social network such as her social links, communities, contents and temporal-spatial behaviors, in a unified manner.

Several other observations are also made from the results: 1) UCGT models perform much better than SVDFeature although they use the same types of features and information, showing the advantage of the well-designed probabilistic generative model incorporating geo-social domain knowledge over the general feature-based matrix factorization model which treats all features equally. 2) UCGT-EG achieves higher prediction accuracy than UCGT-G, showing the benefits of the proposed entropy filtering-based Gibbs sampling. 3) Rank-GeoFM and SVDFeature outperform Geo-SAGE and CBPF in the Foursquare dataset while Geo-SAGE and CBPF exceed Rank-GeoFM and SVDFeature in the Douban Event dataset. The performance disparity may lie in that the location-based social network (e.g., Foursquare) and event-based social network (e.g., Douban Event) have different characteristics. For example, the temporal effect on users' check-in activities is very obvious and thus plays an important role in improving user check-in prediction in LBSNs [8], [35]; while this temporal effect becomes weak in EBSNs, and the content information of events becomes more important.

Parameter Sensitivity Analysis. Tuning model parameters, such as the number of communities (C) and the number of topics (Z) is critical to the performance of our UCGT models. We therefore study the effect of tuning model parameters. As for the hyperparameters α, β, γ and η, for simplicity, we take a fixed value, i.e., $\alpha = 50/Z$, $\gamma = 50/C$ and $\beta = \eta = 0.01$,



(a) On Foursquare Dataset (b) On Douban Event Dataset

Fig. 4. User Check-in Behavior Prediction Accuracy.

following the studies [33], [10]. We try different setups and find that the performance of UCGT models is not sensitive to these hyperparameters. We test the performance of UCGT-EG model by varying the number of topics and communities, and present the results in Tables IV and V. From the results on the Foursquare dataset, we observe that the prediction accuracy first increases with the increasing number of communities, and then it does not change significantly when the number of community is larger than 40. Similar observation is made for increasing the number of topics (i.e., Z): the prediction accuracy increases with the increasing number of topics, and then it does not change much when the number of topics is larger than 30. The reason is that C and Z represent the model complexity. When C and Z are too small, the model has limited ability to describe the data. On the other hand, when C and Z exceed a threshold, the model is expressive enough to handle the data. At this point, it is less helpful to improve the model performance by increasing C and Z. It should be noted that the performance reported in Figure 4(a) is achieved with 40 latent communities (i.e., $C = 40$) and 30 latent topics (i.e., $Z = 30$). Similar observations are also made on the Douban Event dataset, and the experimental results presented in Figure 4(b) are obtained with the parameter settings $C = 200$ and $Z = 150$.

C. Performance in Social Interaction Prediction

We compare our proposed UCGT models with several latest competitors in the application of link prediction. Table III lists the characters of these methods.

Poisson Mixed-Topic Link Model (PMTLM). PMTLM [37] defines a generative process for both text and links between users. Text generation follows the LDA model, and links are modeled as a Poisson distribution. In PMTLM, links and text are generated by the same latent factor, which means one community is bounded to one topic.

Community Level Diffusion Model (COLD). COLD [10] models both topics and communities in a unified latent framework, and extracts inter-community influence dynamics.

Entropy-Based Model (EBM). Pham et al. [21] proposed an entropy-based model to infer social connections and estimate the strength of social connections by analyzing people's co-occurrences in space and time.

TABLE IV
PREDICTION ACCURACY@10 ON FOURSQUARE DATASET.

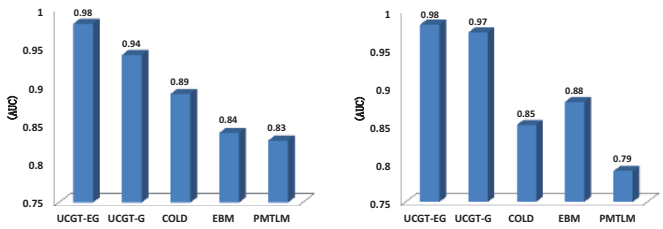
Z \ C	C=10	C=20	C=30	C=40	C=50	C=60
Z=10	0.187	0.205	0.216	0.223	0.223	0.223
Z=20	0.219	0.240	0.253	0.261	0.261	0.261
Z=30	0.228	0.250	0.264	**0.272**	0.272	0.272
Z=40	0.228	0.250	0.264	0.272	0.272	0.273
Z=50	0.229	0.250	0.264	0.272	0.272	0.273

TABLE V
PREDICTION ACCURACY@10 ON DOUBAN EVENT DATASET.

Z \ C	C=50	C=100	C=150	C=200	C=250	C=300
Z=30	0.219	0.240	0.253	0.261	0.261	0.261
Z=50	0.243	0.266	0.280	0.289	0.289	0.289
Z=100	0.253	0.277	0.292	0.301	0.301	0.302
Z=150	0.261	0.286	0.301	**0.311**	0.311	0.311
Z=200	0.261	0.286	0.301	0.311	0.311	0.312
Z=250	0.261	0.286	0.302	0.311	0.312	0.312

Evaluation Method. Since most link prediction methods aim to estimate the probability of a link between two users whereas there is no pre-defined threshold for link existence, we turn to *area under the receiver operating characteristic curve* (AUC) as the prediction accuracy. Given a rank of all non-observed links, the AUC value can be interpreted as the probability that a randomly chosen true positive link is ranked above a randomly chosen true negative link. We adopt a standard 5-fold cross validation, and each time we use 20% of the positive links and randomly select 1% of the negative links to evaluate AUC. The remaining links and all check-in records are used to train the model. As the ground truth of communities is rarely available on social networks, the evaluation of link prediction has been widely used as the proxy of the quantitative measurement of the models developed for community discovery, especially in the mixed-membership community setting without community labels [10].

Experimental Results. Figure 5 shows the AUC values for the five models. Our proposed UCGT models outperform all other methods consistently on the two datasets. This is because UCGT models have the comprehensive modeling ability to exploit all the information of a geo-social network such as



(a) On Foursquare Dataset (b) On Douban Event Dataset

Fig. 5. User Social Interaction Prediction Accuracy.

links, contents and users' temporal-spatial behaviors, and can effectively capture the network proximity, spatiotemporal co-occurrences and semantic similarity among social actors in a unified manner. In contrast, all other network models utilize only portions of the available social networking information, as shown in Table III. Another observation is that the benefit brought by our proposed entropy-based filtering on the Douban Event dataset is not as obvious as on the Foursquare dataset, since most events in the Douban Event dataset are enjoyed by a small group of users with some special interests (e.g., photographing, hiking and IT technology) while most POIs in the Foursquare dataset provide functions or services for the general public (e.g., shopping centers, hospitals and cinemas).

Parameter Sensitivity Analysis. We also study the impact of varying parameters in UCGT-EG, e.g., the number of communities (C) and the number of topics (Z), and present the results in Tables VI and VII. From the results, we make the similar observation to what is found in Tables IV and V: the link prediction accuracy of UCGT first increases with the increasing numbers of communities and topics, and then it does not change much when the numbers of communities and topics are larger than a threshold.

TABLE VI
LINK PREDICTION ACCURACY (AUC) ON FOURSQUARE DATASET.

Z \ C	C=10	C=20	C=30	C=40	C=50	C=60
Z=10	0.793	0.848	0.884	0.902	0.902	0.902
Z=20	0.837	0.894	0.932	0.951	0.951	0.951
Z=30	0.862	0.921	0.960	**0.980**	0.980	0.980
Z=40	0.862	0.921	0.960	0.980	0.980	0.980
Z=50	0.862	0.921	0.960	0.980	0.980	0.980

TABLE VII
LINK PREDICTION ACCURACY (AUC) ON DOUBAN EVENT DATASET.

Z \ C	C=50	C=100	C=150	C=200	C=250	C=300
Z=30	0.691	0.757	0.799	0.823	0.823	0.823
Z=50	0.766	0.838	0.884	0.911	0.911	0.911
Z=100	0.799	0.875	0.922	0.951	0.951	0.951
Z=150	0.823	0.902	0.951	**0.980**	0.980	0.980
Z=200	0.823	0.902	0.951	0.980	0.980	0.980
Z=250	0.823	0.902	0.951	0.980	0.980	0.981

D. Model Training Efficiency

In this experiment, we evaluate the efficiency of UCGT model training on the large-scale Douban Event dataset. To tackle the challenge of large data size and ensure the scalability of our UCGT models, we deploy our inference algorithm of UCGT to three parallel computation settings: *Multicore*, *Cluster* and their combination *Multicore-Cluster*. We conduct this experiment on a cluster consisting of 10 servers (Dell R630 Rack Server). Each server is equipped with 2 processors (Intel Xeon E5-v3), 32 cores and 64 GB memory. We run 10 threads in each server node. Besides, we also compare our developed parallel mechanisms with the MapReduce framework implemented by Hadoop on the same cluster. To fairly compare with our methods, we use Memcached instead of HDFS in MapReduce.

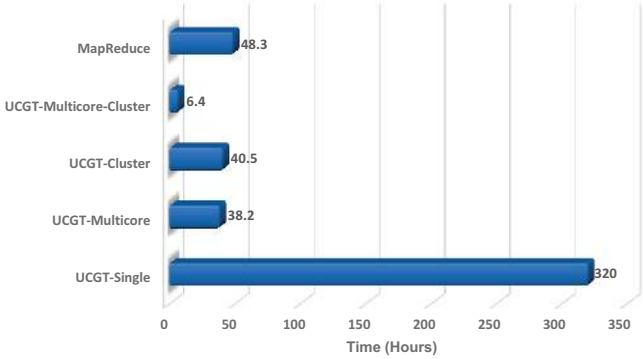


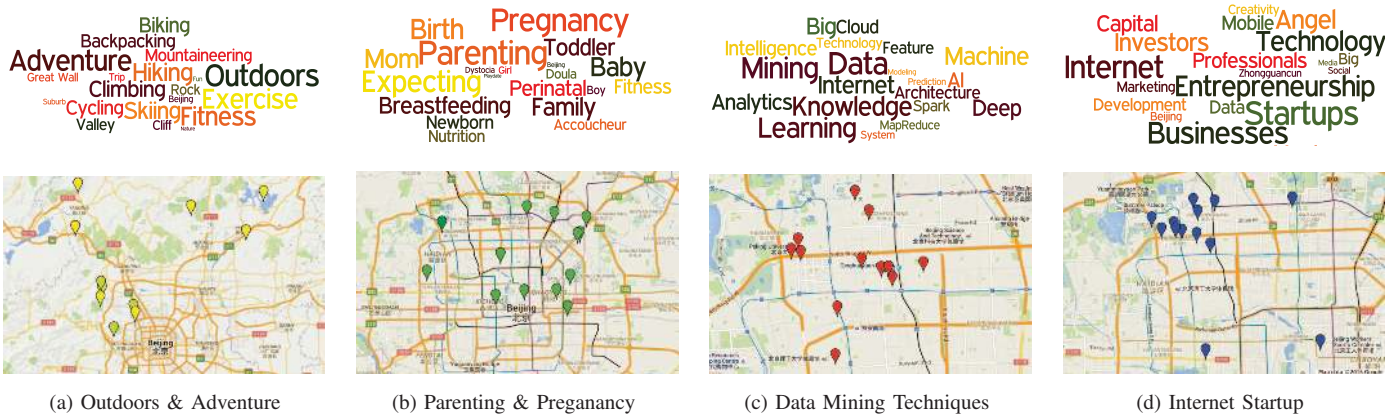
Fig. 6. Model Training Time

Figure 6 shows the running time of different methods on the Douban Event dataset. Although the basic implementation of UCGT (i.e., UCGT-Single in Figure 6) is costly, our parallel implementations guarantee the efficiency. Specifically, we reduce the training time for 18.9M check-in records and 30M links from hundreds of hours (320) to just a few hours (6.4). This clearly shows the advantage of parallel processing by leveraging the power of multicores and clusters. The model structure of UCGT is loosely coupled enough to facilitate parallel processing. Thus, our proposed UCGT models are scalable to large-scale geo-social networking data. Besides, our developed three parallel implementations, especially UCGT-Multicore-Cluster, outperform MapReduce due to the following reasons: 1) MapReduce only takes advantage of the parallelization brought by the cluster and ignores the potential of the multicores. 2) Due to a number of reasons (system, disk access, general job load, sampler burn-in), the time of each node to process the geo-social data may differ widely. Waiting for the last node to finish before synchronization can occur, introduces potentially long idle times in MapReduce.

E. Qualitative Analysis of Detected Communities

In this experiment, we use a case study method to demonstrate the effectiveness of UCGT in detecting communities qualitatively. For an intuitive understanding of the discovered communities, we choose the top-20 words with the highest generation probabilities for each community on the Douban Event dataset. Specifically, given a community c and a word w, the probability of c generating w is computed as follows:

$$P(w|c, \hat{\mathbf{\Psi}}) = \sum_{z} \hat{\theta}_{c,z} \hat{\phi}_{z,w}.$$



(a) Outdoors & Adventure (b) Parenting & Preganancy (c) Data Mining Techniques (d) Internet Startup

Fig. 7. Semantic and Geographical Interpretations of Discovered Communities.

Based on the top words and their corresponding generation probabilities, we create word clouds for each discovered community. We present four example word clouds in Figure 7. Besides, for each community c, we also present the locations of the top-20 events with the highest generation probabilities $\vartheta_{c,v}$ on Google Maps.

From the results, we observe that all these four communities were living in Beijing, China, but they focused on different subjects. For example, the members of the third community were interested in data mining techniques, and they often attended the salons, seminars or lectures related with data mining techniques together. From the corresponding geographical map, we can see that most of these events were held in Zhongguancun, Haidian District Beijing, such as Peking University, Tsinghua University and MSR Asia. Thus, the communities discovered by UCGT are semantically and geographically interpretable. Capturing the communities' interests and spatial ranges of activity are critical to improve recommendations, search and target ads for a group of users.

VI. Related Work

Community Detection. Communities are natural groups formed by users with close connections and similar interests [12]. A mixed membership stochastic block model was introduced in [1] where each user has a probability distribution over communities. A growing number of recent works [10], [37] incorporated both the network structure and content to improve community detection performance. In these models, content and links are both generated by the same latent variables. Thus communities are limited to have one-to-one correspondence with topics. Our model decomposes these two factors, which opens up an array of meaningful and desired extraction such as community interests over topics. Qi et al. [23] proposed to leverage edge content to improve the effectiveness of community detection in email networks. However, the edge content is not available in most of online social networks, especially the geo-social networks. Moreover, to the best of our knowledge, this is the first work to consider spatiotemporal co-occurrence information for community discovery.

Inferring Social Ties from Geographic Coincidences. Using the geographical records to infer users' social behaviors and relationships is a hot topic in spatiotemporal data mining [21], [6], [26]. The methods proposed in [6] have investigated the meeting events that occur at different times (e.g., weekdays vs. weekends or day vs. night) to infer different types of relationships such as friends and colleagues. Meanwhile, Crawnshaw et al. [6] extracted a set of features from both meeting events and the individual mobility patterns and learned a model to identify the friendship from users' check-in data. Pham et al. [21], [26] further considered the diversity of meeting locations to handle cases that two users meet by coincidence. However, all these methods have not considered the semantic information of meeting events, nor the existing social network structure, thus they cannot interpret why the meeting events occur.

Check-in Behavior Prediction. Many recent studies [29], [5] showed that there is a strong correlation between user check-in activities and geographical distance as well as social connections, thus most existing check-in behavior prediction (or location recommendation) work mainly focuses on leveraging the *geographical and social influences* to improve prediction accuracy. For example, Ye et al. [29] delved into POI recommendation by investigating the geographical influences among locations and proposed a framework that combines user preferences, social influence and geographical influence. Cheng et al. [4] investigated the geographical influence through combining a multi-center Gaussian model, matrix factorization and social influence together for location prediction. Lian et al. [14] incorporated spatial clustering phenomenon resulted by geographical influence into a weighted matrix factorization framework to deal with the challenge from matrix sparsity. The *temporal effect* of user check-in activities in LBSNs has also attracted much attention from researchers. The prediction methods with temporal effect mainly leverage temporal cyclic patterns and temporal chronological patterns on LBSNs [8], [35]. Yin et al. [32] was the first to study the problem of real-time POI recommendation. Most recently, researchers explored the *content information* of spatial items to alleviate the problem of data sparsity, especially in the out-of-town recommendation [27], [34], [30], [31]. Compared with our UCGT models, these existing models utilize only portions

or a few aspects of geo-social network information and lack a comprehensive modeling ability. Besides, our UCGT makes most of the community members' collective behavior patterns to overcome the sparsity of individual's check-in behaviors.

VII. Conclusion

In this paper, we studied how to discover interpretable communities from a geo-social network by capturing all its information such as social links, semantic contents, spatial and temporal information in a unified manner. Technically, we proposed a Bayesian model UCGT to define the generative process of communities as a result of network proximities, spatiotemporal co-occurrences and semantic similarity. To improve the performance of UCGT, we incorporated the idea of entropy filtering to Gibbs Sampling. To adapt to large-scale geo-social data, we developed a scalable parallel implementation of the UCGT by harnessing the powers of multicores and clusters. UCGT can also be used to address various practical problems, such as check-in prediction and link prediction. We evaluated the performance of UCGT on two large-scale geo-social networking datasets, and the experimental results demonstrated its superiority in terms of prediction accuracy, modeling training efficiency and community interpretability.

Acknowledgement

This work is partially supported by ARC Discovery Early Career Researcher Award (DE160100308), National Basic Research Program of China (2013CB329305), ARC Discovery Project (DP140103171 and DP120102829). It is also partially supported by National Natural Science Foundation of China (Grant No. 61572335, 61502466, 61232006, 61303164, 61402447), Beijing Natural Science Foundation (Grant No. 9144037) and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

References

[1] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.*, 9:1981–2014, June 2008.

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

[3] T. Chen, W. Zhang, Q. Lu, K. Chen, Z. Zheng, and Y. Yu. Svdfeature: A toolkit for feature-based collaborative filtering. *J. Mach. Learn. Res.*, 13(1):3619–3622, Dec. 2012.

[4] C. Cheng, H. Yang, I. King, and M. R. Lyu. Fused matrix factorization with geographical and social influence in location-based social networks. In *AAAI*, 2012.

[5] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *KDD*, pages 1082–1090, 2011.

[6] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh. Bridging the gap between physical location and online social networks. In *UbiComp*, pages 119–128, 2010.

[7] N. Eagle, A. S. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *PNAS*, 106(36):15274–15278, 2009.

[8] H. Gao, J. Tang, X. Hu, and H. Liu. Exploring temporal effects for location recommendation on location-based social networks. In *RecSys*, pages 93–100, 2013.

[9] Q. Ho, R. Yan, R. Raina, and E. P. Xing. Understanding the interaction between interests, conversations and friendships in facebook. *arXiv preprint arXiv:1211.0028*, 2012.

[10] Z. Hu, J. Yao, B. Cui, and E. Xing. Community level diffusion extraction. In *SIGMOD*, pages 1555–1569, 2015.

[11] G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM J. Sci. Comput.*, 20(1):359–392, Dec. 1998.

[12] J. Leskovec, K. J. Lang, and M. Mahoney. Empirical comparison of algorithms for network community detection. In *WWW*, pages 631–640, 2010.

[13] X. Li, G. Cong, X.-L. Li, T.-A. N. Pham, and S. Krishnaswamy. Rank-geofm: A ranking based geographical factorization method for point of interest recommendation. In *SIGIR*, pages 433–442, 2015.

[14] D. Lian, C. Zhao, X. Xie, G. Sun, E. Chen, and Y. Rui. Geomf: Joint geographical modeling and matrix factorization for point-of-interest recommendation. In *KDD*, pages 831–840, 2014.

[15] M. Lichman and P. Smyth. Modeling human location data with mixtures of kernel densities. In *KDD*, pages 35–44, 2014.

[16] X. Liu, Q. He, Y. Tian, W.-C. Lee, J. McPherson, and J. Han. Event-based social networks: linking the online and offline social worlds. In *KDD*, pages 1032–1040, 2012.

[17] U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, Dec. 2007.

[18] H. Ma, H. Yang, M. R. Lyu, and I. King. Sorec: Social recommendation using probabilistic matrix factorization. In *CIKM*, pages 931–940, 2008.

[19] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[20] D. Newman, A. Asuncion, P. Smyth, and M. Welling. Distributed algorithms for topic models. *The Journal of Machine Learning Research*, 10:1801–1828, 2009.

[21] H. Pham, C. Shahabi, and Y. Liu. Ebm: An entropy-based model to infer social strength from spatiotemporal data. In *SIGMOD*, pages 265–276, 2013.

[22] T.-A. N. Pham, X. Li, G. Cong, and Z. Zhang. A general graph-based model for recommendation in event-based social networks. In *ICDE*, pages 567–578, 2015.

[23] G.-J. Qi, C. Aggarwal, and T. Huang. Community detection with edge content in social media networks. In *ICDE*, pages 534–545, 2012.

[24] P. Sarkar and A. W. Moore. Dynamic social network analysis using latent space models. *SIGKDD Explor. Newsl.*, 7(2):31–40, Dec. 2005.

[25] A. Smola and S. Narayanamurthy. An architecture for parallel topic models. *Proc. VLDB Endow.*, 3(1-2):703–710, Sept. 2010.

[26] H. Wang, Z. Li, and W.-C. Lee. Pgt: Measuring mobility relationship using personal, global and temporal factors. In *ICDM*, pages 570–579, Dec 2014.

[27] W. Wang, H. Yin, L. Chen, Y. Sun, S. Sadiq, and X. Zhou. Geo-sage: A geographical sparse additive generative model for spatial item recommendation. In *KDD*, pages 1255–1264, 2015.

[28] J. Xie, S. Kelley, and B. K. Szymanski. Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Comput. Surv.*, 45(4):43:1–43:35, Aug. 2013.

[29] M. Ye, P. Yin, W.-C. Lee, and D.-L. Lee. Exploiting geographical influence for collaborative point-of-interest recommendation. In *SIGIR*, pages 325–334, 2011.

[30] H. Yin, B. Cui, Z. Huang, W. Wang, X. Wu, and X. Zhou. Joint modeling of users' interests and mobility patterns for point-of-interest recommendation. In *SIGMM*, pages 819–822, 2015.

[31] H. Yin, B. Cui, Y. Sun, Z. Hu, and L. Chen. Lcars: A spatial item recommender system. *ACM Trans. Inf. Syst.*, 32(3):11:1–11:37, July 2014.

[32] H. Yin, B. Cui, X. Zhou, W. Wang, Z. Huang, and S. Sadiq. Joint modeling of user check-in behaviors for real-time point-of-interest recommendation. *ACM Trans. Inf. Syst.*, 2016.

[33] H. Yin, Y. Sun, B. Cui, Z. Hu, and L. Chen. Lcars: A location-content-aware recommender system. In *KDD*, pages 221–229, 2013.

[34] H. Yin, X. Zhou, Y. Shao, H. Wang, and S. Sadiq. Joint modeling of user check-in behaviors for point-of-interest recommendation. In *CIKM*, pages 1631–1640, 2015.

[35] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. M. Thalmann. Time-aware point-of-interest recommendation. In *SIGIR*, pages 363–372, 2013.

[36] W. Zhang and J. Wang. A collective bayesian poisson factorization model for cold-start local event recommendation. In *KDD*, pages 1455–1464, 2015.

[37] Y. Zhu, X. Yan, L. Getoor, and C. Moore. Scalable text and link analysis with mixed-topic link models. In *KDD*, pages 473–481, 2013.