frontiers in Big Data

ORIGINAL RESEARCH
published: 10 June 2021
doi: 10.3389/fdata.2021.690970

Measuring Urban Vibrancy of Residential Communities Using Big Crowdsourced Geotagged Data

*Pengyang Wang, Kunpeng Liu, Dongjie Wang and Yanjie Fu**

Computer Science Department, University of Central Florida, Orlando, FL, United States

The pervasiveness of mobile and sensing technologies today has facilitated the creation of Big Crowdsourced Geotagged Data (BCGD) from individual users in real time and at different locations in the city. Such ubiquitous user-generated data allow us to infer various patterns of human behavior, which helps us understand the interactions between humans and cities. In this article, we aim to analyze BCGD, including mobile consumption check-ins, urban geography data, and human mobility data, to learn a model that can unveil the impact of urban geography and human mobility on the vibrancy of residential communities. Vibrant communities are defined as places that show diverse and frequent consumer activities. To effectively identify such vibrant communities, we propose a supervised data mining system to learn and mimic the unique spatial configuration patterns and social interaction patterns of vibrant communities using urban geography and human mobility data. Specifically, to prepare the benchmark vibrancy scores of communities for training, we first propose a fused scoring method by fusing the frequency and the diversity of consumer activities using mobile check-in data. Besides, we define and extract the features of spatial configuration and social interaction for each community by mining urban geography and human mobility data. In addition, we strategically combine a pairwise ranking objective with a sparsity regularization to learn a predictor of community vibrancy. And we develop an effective solution for the optimization problem. Finally, our experiment is instantiated on BCGD including real estate, point of interests, taxi and bus GPS trajectories, and mobile check-ins in Beijing. The experimental results demonstrate the competitive performances of both the extracted features and the proposed model. Our results suggest that a structurally diverse community usually shows higher social interaction and better business performance, and incompatible land uses may decrease the vibrancy of a community. Our studies demonstrate the potential of how to best make use of BCGD to create local economic matrices and sustain urban vibrancy in a fast, cheap, and meaningful way.

Keywords: urban vibrancy, spatiotemporal data mining, urban computing, Big Crowdsourced Geotagged Data mining, learning-to- rank

OPEN ACCESS

Edited by:
Xun Zhou,
The University of Iowa, United States

Reviewed by:
Zhe Jiang,
University of Alabama, United States
Tong Xu,
University of Science and Technology of China, China

****Correspondence:***
Yanjie Fu
yanjie.fu@ucf.edu

Specialty section:
This article was submitted to Data Mining and Management, a section of the journal Frontiers in Big Data

Received: *05 April 2021*
Accepted: *12 May 2021*
Published: *10 June 2021*

Citation:
Wang P, Liu K, Wang D and Fu Y (2021) Measuring Urban Vibrancy of Residential Communities Using Big Crowdsourced Geotagged Data. Front. Big Data 4:690970. doi: 10.3389/fdata.2021.690970

INTRODUCTION

Vibrant residential communities (vibrant communities for short) are defined as places that show diverse and frequent consumer activities. Vibrant communities usually have the following features: permeability, vitality, variety, accessibility, identity, and legibility. Developing vibrant communities is very beneficial for both social good and business good. For instance, vibrant communities can attract talented younger workers, high-tech entrepreneurs, and cutting-edge firms, as well as foster intensive social interactions, productivity, and creative activities. Thereby, understanding urban vibrancy can help 1) contribute to economic growth; 2) enhance public security; and 3) improve environmental, fiscal, and social outcomes. For example, when hunting for a business site, entrepreneurs should consider the surrounding community, whether it is welcoming and attractive for business activities (Church and Murray, 2009). By studying the urban vibrancy patterns of communities, we can make better decisions and suggestion for business site selection, to ensure successful business.

However, it is traditionally challenging to develop vibrant communities because there is not a clear answer to the following question: "what kind of communities tend to have higher vibrancy?" In prior literature, researchers have conducted conceptual and empirical studies about vibrant communities in the fields of urban planning and social science. For example, Glaeser *et al.* pointed out that vibrant communities depend on the demand for urban *density* (Glaeser et al., 2001). Couture *et al.* found people are willing to pay higher rents and transportation costs for vibrant places (Couture, 2013). Farber *et al.* found that vibrant communities are associated with spatial concentration of residents and diversity of products and services (Farber and Li, 2013). Malizia *et al.* found that vibrant communities are usually compact, dense, and accessible with diverse land uses (Malizia and Song, 2014). Neutens *et al.* found that high-density and mixed land uses can benefit quality social interactions and enhance community vibrancy (Neutens et al., 2013). Dougal *et al.* argued urban vibrancy can be reflected by dynamic human-dependent factors (e.g., highly talented workers) that vary over time (Dougal et al., 2015). However, all these studies only provide conceptual understanding on one or two aspects of the community vibrancy.

In order to provide a comprehensive understanding of various aspects that contribute to the community vibrancy, we propose a big data–driven approach which is the first time to systematically study the measurements and patterns of vibrant communities. Specifically, we take advantage of the large-volume and ubiquitous user-generated data collected from diverse sources, for example, buildings, vehicles, human, sensors, and devices, in real time and at different locations in the city. Such Big Crowdsourced Geotagged Data (BCGD) allow us to infer various patterns of human behavior and understand the interactions between humans and cities. If properly analyzed, these data can be a rich source of intelligence to discover and mimic the unique spatial and mobility patterns of vibrant communities.

However, due to the variety and veracity nature of big data, it is very challenging to analyze BCGD. To make the analysis effective and efficient, we propose to focus the community vibrancy analysis on two perspectives: 1) spatial configuration and 2) social interaction. First, the spatial configuration of a community is empirically defined as the physical characteristics that make up built-up areas, such as bus systems, subway systems, road networks, and landmarks, as well as corresponding locations, numbers, and mutual distances. Prior literature has developed empirical evidence that suggests the significant impact of spatial configuration on community vibrancy (Song and Knaap, 2004; Koster and Rouwendal, 2012; Loehr, 2013). However, it is not a trivial task to quantify the spatial configuration of communities. Particularly, we need to construct effective variables (i.e., features) from static urban geography data (e.g., landmarks, public transportation data, and road network data), in order to capture the compatible dimensions of spatial structure, as well as the corresponding portfolios and geographic allocations of these dimensions within a community. Second, from the perspective of social interaction, there are some preliminary studies (Farber et al., 2013, 2014; Farber and Li, 2013; Neutens et al., 2013) about measuring general social interactions using human mobility data. Unfortunately, since human mobility data are mostly in a form of trajectories or footprints, typically represented by a sequence of GPS location points, such data are lack of semantically rich information, which makes the task of profiling social interactions within and across communities very challenging. Therefore, we propose to augment and enrich the semantic information of human mobility data in order to analyze intercommunity and intra-community social interaction. In summary, we propose to analyze and extract the features of spatial configuration from urban geography data and the features of social interaction from human mobility to spot highly vibrant communities, which will be formulated as a ranking-based data mining task next.

Although a lot of features may be extracted from a variety of data sources, these extracted vibrancy-related features are often correlated and redundant. The feature redundancy can result in poor generalization performance. In reality, a small number of good features are usually sufficient to represent the patterns of vibrant communities and facilitate accurate prediction of spotting vibrant communities. Conventional methods usually use a two-step paradigm, which is basically to first select a feature subset and then learn a ranking model based on the selected features. However, the selected feature subset may not be optimal for ranking because the two steps are modeled separately. As revealed by many machine learning researchers, the presumption of the sparsity-regularized classification models is that only a subset of features are significant for prediction; that is, the coefficients of nonsignificant features will be very small and close to zero in the learned classification model. Therefore, we propose to combine sparsity regularization and ranking objective in a unified model to help us identify the optimal feature subset for spotting vibrant communities.

To summarize, in this article, we conduct a systematic study on the measurements, patterns, and modeling of urban vibrancy.

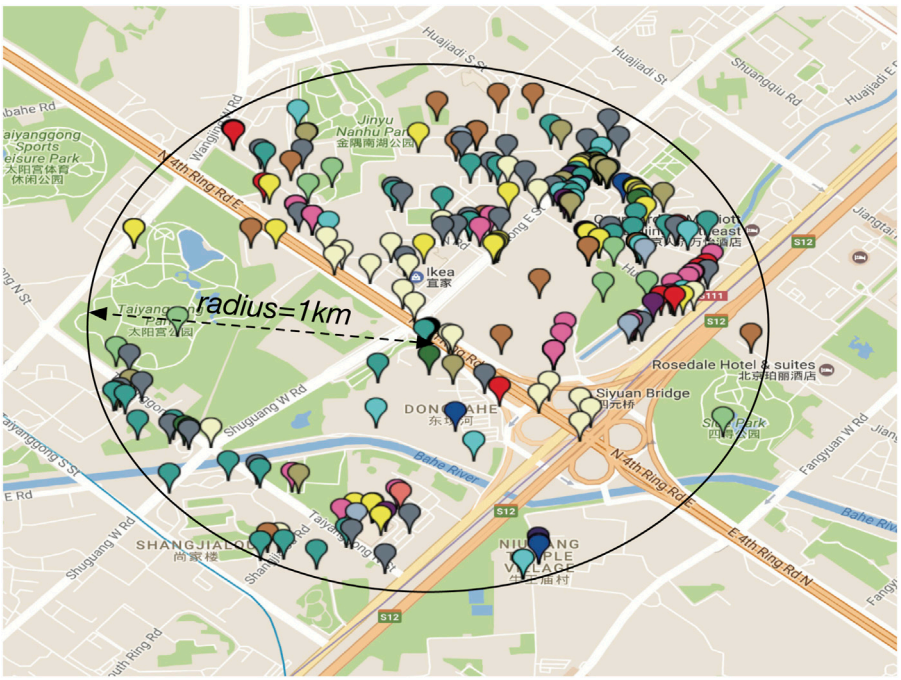

FIGURE 1 | Example of a residential community.

Specifically, the following are our main contributions: 1) We start with defining a fused scoring method based on F-measure to quantify the urban vibrancy of communities. 2) We mine the features of spatial configuration from static urban geography data and the features of social interactions from dynamic human mobility data. 3) Given the obtained features, we develop a novel model to learn the patterns of vibrant communities, by combining pairwise ranking objective and sparsity regularization in a unified probabilistic framework, which is greatly enhanced by simultaneously conducting feature selection and maximizing ranking accuracy. 4) Finally, we conduct comprehensive performance evaluations for the feature sets and models with large-scale real-world data, and the experimental results demonstrate the competitive performance of our method with respect to different validation metrics.

PROBLEM STATEMENTS AND FRAMEWORK OVERVIEW

In this section, we first introduce the important definitions and formulate the problem. After that, we provide an overview of the proposed analytic framework.

Definitions and Problem Formulation

Residential community: A residential community consists of a location (i.e., latitude and longitude) of a residential complex and a neighborhood area (e.g., a circle with radius of 1 km). A residential complex often includes one or multiple apartment buildings in urban areas. There are a variety of point of interests (POIs) in the neighborhood area, providing many services to people. **Figure 1** shows an example of a residential community.

Problem definition: Formally, given a set of I residential communities $X = \{x_1, x_2, \ldots, x_I\}$, the goal of our problem is to rank them in a descending order according to their vibrancy scores $Y = \{y_1, y_2, \ldots, y_I\}$. BCGD such as point of interest data, human mobility data, and mobile consumption check-in data have encoded the unique spatial and social patterns of residential communities, and thus can be used to identify vibrant communities by exploiting a data-driven analytics-enabled strategy. Essentially, there are three major tasks: 1) Developing empirical and measurable metric to quantify the vibrancy scores of residential communities; 2) quantifying the patterns of spatial configuration and social interaction within and across residential communities; and 3) learning to spot highly vibrant communities with the spatial and social patterns of communities.

Framework Overview

The focus of this article is to develop a data mining approach for spotting vibrant residential communities. In the pursuit of this general aim, we have three specific tasks: measurement, patterns, and modeling.

- In researching measurements, we aim to develop an empirical metric to measure community vibrancy using a data-driven strategy. While urban vibrancy is difficult to be observed, BCGD provide a potential to circumvent this problem. To quantify vibrancy empirically, we make use of novel mobile consumption check-in data and propose an unsupervised fused scoring method to quantify the vibrancy score of each community.

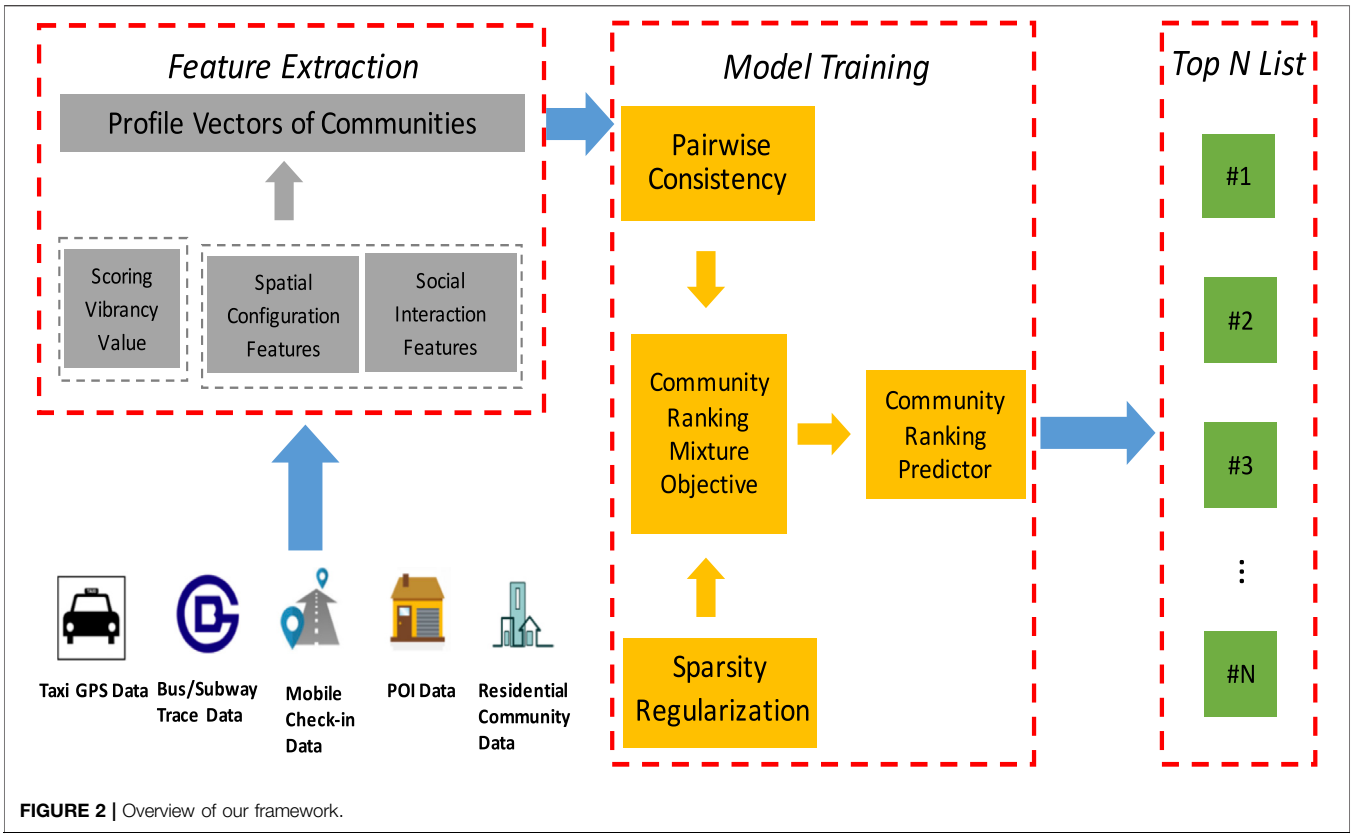

FIGURE 2 | Overview of our framework.

- In researching patterns, we aim to discover the patterns of community vibrancy. We extract various contextual features from two perspectives: spatial configuration and social interactions. The spatial configuration features are extracted from the urban geography data including public transportation, road networks, and POIs; the social interaction features are extracted from the human mobility data including bus GPS data, taxi GPS data, and smartphone GPS data.
- In researching patterns, to make full use of all relevant features, we develop a sparse learning-to-rank approach for spotting vibrant communities.

Figure 2 shows the overview of the proposed analytic framework.

AN EMPIRICAL METRIC FOR ESTIMATING COMMUNITY VIBRANCY

In prior literature, researchers have found that the vibrancy score of a residential community can be reflected by consumer activities from two perspectives: density and diversity of consumer activities (Talen, 1999; Glaeser et al., 2001; Couture, 2013; Farber and Li, 2013; Malizia and Song, 2014). Here, "density" can be explained by the fact that if a large number of consumers are willing to pay higher transportation costs to visit a place, and to spend more time to consume in that place, the place is likely to be vibrant. A high "diversity" of consumer activities indicates that this place can meet a variety of consumption needs and help consumers carry out different outdoor activities in a single place within a walking distance. In other words, consumers do not have to visit other places and can complete a variety of activities in a single place.

To capture vibrancy empirically, we make use of a novel geotagged user consumption check-in data shared in location-based social networks (LBSNs). A check-in event contains the information of a mobile user's destination POI and consumption activity type, which connects user profiles, POI locations, and outdoor activities with measurable density and diversity. The presumption is that urban vibrancy increases the density and diversity of consumer activities and POIs in a place. In other words, urban vibrancy promotes the probability that mobile users check into a place, enhances the diversity of urban functions, and improves social interactions and centralization across different categories of outdoor activities. With this presumption, urban vibrancy, even though not observed directly, can be identified by strategically fusing the observable densities and diversities of mobile check-ins over various activity categories, for example, home, work, date, dinning, travel, transportation, shopping, and entertainment. Specifically, urban vibrancy can be quantified by mathematically giving a vibrancy score using a fused scoring framework. We propose to proceed with three steps: 1) measuring the density of consumer activities, 2) measuring the diversity of consumer activities, and 3) fused scoring.

1) Measuring the density of consumer activities.

We propose to extract the density of consumer activities in communities. For each residential community, we count the total number of mobile consumption check-in events (#) as an estimation of the density of consumer activities, denoted by fre = #.

2) Measuring the diversity of consumer activities.

To estimate the diversity of consumer activities in a community, we count the numbers of mobile check-in events with respect to different POI categories, denoted by $\{\#_c\}_{c=1}^{\mathbb{C}}$, where c represents the c-th POI category and $\mathbb{C}$ denotes the number of POI categories. We compute the diversity of consumer activities by exploiting the definition of entropy

$$\text{div} = \sum_{c=1}^{\mathbb{C}} \#_c \log \#_c. \tag{1}$$

3) Fused scoring.

After extracting and normalizing both density and diversity, we use the F-1 score

$$\text{vibrancy} = \frac{2 * \text{fre} * \text{div}}{\text{fre} + \text{div}} \tag{2}$$

to fuse both density and diversity into a single score. The score extracted from consumer activities can empirically measure the vibrancy of a community.

DISCOVERING PATTERNS OF VIBRANT COMMUNITIES

We now proceed to introduce discriminative features to describe and quantify the patterns of vibrant communities. Specifically, we categorize the features into two categories:

- *The features of spatial configurations*, which can be extracted from urban geography data, such as public transportation, road networks, and POIs.
- *The features of social interactions*, which can be extracted from human mobility data (taxi GPS data, bus GPS data, etc.) within and across communities.

Features of Spatial Configuration

The spatial configuration of a community is a three-element tuple, including 1) the compatible dimensions of the spatial configuration, such as shopping, living service, education, and transportation buildings that serve important urban functions; 2) the portfolio of these compatible dimensions, such as frequency, density, and diversity of different POIs; and 3) the geographic allocation of these compatible dimensions, such as distances to different POIs. The recent study by Evans et al. (2007) implied that urban environmental elements combine to determine the quality of life in higher density and mixed-use locations. Moreover, the study by Yue et al. (2017) showed that POI diversity contributes significantly to improving neighborhood vibrancy. Therefore, we extract 1) density of POIs, 2) diversity of POIs, and 3) accessibility of transportation as features for each community c_i, in which there are a set of POIs, denoted by $P = \{p : p \in c_i \& p \in P\}$, where p is a POI.

1) Density of POIs.

After studying large-scale residential community data, mobile check-in data, and POI data, we found that the vibrancy level of a place depends on the density of POIs in the same area. Intuitively, the more POIs in a community, the more likely the community could meet a visitor's various needs, such as dating, shopping, and watching movies. Therefore, we exploit the density of POIs as a feature. Specifically, for each community c_i, we can count the POI number of each POI category ϕ_k. POI categories are defined based on their functions, such as shopping, sports, and education. Formally, we have

$$num_{c_i}^{\phi_k} = \sum I\{p \in c_i \& p \in P \& p \in \phi_k\}, \tag{3}$$

where I denotes the numbers of POIs. We note that given that the radius of a community is the same, the density depends only on the number of POIs they include.

2) Diversity of POIs.

To assess the influence of the spatial heterogeneity of community functionalities on the vibrancy of a community, we apply the entropy measure to describe the diversity of POIs for a community. For each community c_i, we calculate the diversity η_i as follows:

$$\eta_i = -\sum_k \frac{num_{c_i}^{\phi_k}}{\sum_k num_{c_i}^{\phi_k}} \log \frac{num_{c_i}^{\phi_k}}{\sum_k num_{c_i}^{\phi_k}}. \tag{4}$$

According to the definition, the larger the entropy is, the higher diversity the community has. Be sure to notice that the diversity of POIs has correlation with the diversity of user consumption activities in the measurement section of community vibrancy, but they are two different concepts. The diversity of POIs represents the spatial configuration and geographic allocation of a community; the diversity of user consumption activities denotes a quantitative aspect of human dynamic behavior.

3) Accessibility.

We refer accessibility to the degree of convenience that consumers can visit a community. For example, street connectivity, higher bus stop density, and greater nonmotorized access promote the possibility of human mobility and influence the transportation mode choice (Khan et al., 2014); different effects of spatial accessibility vary among different trip purposes (Zhang, 2005); and users in different

gender and youth groups show different mobility patterns in rural and suburban areas (Collins et al., 2012). Generally, public transportation facilities and the quality of the road network are two basic factors that influence the accessibility.

$$num_{c_i}^{b} = \sum I\{b \in B \& b \in c_i\}, \tag{5}$$

where I denotes the numbers of bus stops, B denotes the bus stop set, and b denotes a bus stop. Besides, we also calculate the minimum distance from POIs to the bus stops as $\zeta_{c_i}^{b}$:

$$\zeta_{c_i}^{b} = \min_{p \in c_i \& b \in c_i} dist(p, b), \tag{6}$$

where $dist(p, b)$ denotes the distance between a POI p and a bus station b. Similarly, for subway stations, we calculate the number of subway stations:

$$num_{c_i}^{s} = \sum I\{s \in S \& s \in c_i\}, \tag{7}$$

where S denotes the subway station set and s denotes a subway station. And the minimum distance from POIs to the subway stations is denoted as $\zeta_{c_i}^{s}$:

$$\zeta_{c_i}^{s} = \min_{p \in c_i \& s \in c_i} dist(p, s), \tag{8}$$

where $dist(p, s)$ denotes the distance between a POI p and a subway station s.

1. Public transportation facilities. There are two major types of public transportation—bus and subway, in most big cities. Therefore, we define and extract some important properties of *bus stops* and *subway stations*. For each community c_i, we calculate the number of bus stations:
2. The quality of the road network. Intuitively, in an urban area, if a community has more intersections of road networks, consumers can access taxis or enter road network systems by private cars more easily. Also, if a community with the same radius has longer roads and highways, the density of road networks is higher. Therefore, we calculate *the number of intersections of roads* (denoted as $num_{c_i}^{\tau}$) and *the density of road networks* (denoted as v_i) to measure the quality of road networks.

For each community c_i, the number of intersections of roads $num_{c_i}^{\tau}$ can be calculated as

$$num_{c_i}^{\tau} = \sum_{\tau} I\tau \in c_i, \tag{9}$$

where τ denotes an intersection in the road networks, and the density of road networks v_i can be calculated as

$$v_i = \sum_{\tau_k \in c_i \& \tau_l \in c_i} dist(\tau_k, \tau_l), \tag{10}$$

where $dist(\tau_k, \tau_l)$ denotes the distance between two intersections τ_k and τ_l.

Features of Social Interactions

Social interactions within and across communities can be observed and estimated from people's movements. In general, human mobility encodes two types of social interactions:

- *The interactions between users and users*: A mobile user moves from one community to another community and stays in the destination for a certain time. During this time period, the mobile user is highly likely to meet and speak to other mobile users, particularly for the trip purpose of dating, entertainment, and dinning.
- *The interactions between users and places*: Mobile users inevitably have to interact with a variety of POIs to complete activities with respect to different trip purposes, such as working, shopping, dining, and entertainment.

As a result, we extract social interaction features from the human mobility data based on the following three perspectives: *(i) mobility flow, (ii) range, and (iii) average speed.*

1) Mobility flow.

Taking a community as an example, we can observe movements that leave a community, arrive at a community, and transit within a community. Based on the above observations, all movements can be segmented into three types: *1) inflow* (corresponding to arriving human mobility), *2) outflow* (corresponding to leaving human mobility), and *3) intra-flow* (corresponding to human mobility within communities). In BCGD, a movement trajectory tr_k can be represented as a four-element tuple $(O_k, t_{O_k}, D_k, t_{D_k})$, where O_k denotes the original point, t_{O_k} denotes the start time, D_k denotes the destination point, and t_{D_k} denotes the end time.

$$inflow_{c_i} = \sum_{k} I\{O_k \notin c_i, D_k \in c_i\}, \tag{11}$$

where I denotes the number of trajectories and $inflow_{c_i}$ denotes the inflow volume of the community c_i.

$$outflow_{c_i} = \sum_{i} I\{O_k \in c_i, D_k \notin c_i\}, \tag{12}$$

where I denotes the numbers of trajectories and $outflow_{c_i}$ denotes the outflow volume of the community c_i.

$$intra - flow_{c_i} = \sum_{i} I\{O_k \in c_i, D_k \in c_i\}, \tag{13}$$

where I denotes the number of trajectories and $intra - flow_{c_i}$ denotes the intra-flow volume of the community c_i.

2) Range.

1. Inflow interaction. Inflow is defined as movements that people come to visit the community c_i from other communities. Therefore, the volume of inflow can be calculated as
2. Outflow interaction. Outflow is defined as movements that people leave the community c_i. Therefore, the volume of outflow can be calculated as
3. Intra-flow interaction. Intra-flow is defined as movements that are inside of the community c_i. Therefore, the volume of intra-flow can be calculated as

We check the maximum commute distance of taxis to the community to represent the range of social interactions. For a community c_i, we calculate the range of interaction λ_i as

TABLE 1 | Feature summary.

Feature type	Category	Subcategory	Denotation
Spatial configuration	Density		$num_{c_i}^{\phi_k}$
	Diversity		η_i
	Accessibility	Public transportation facilities	$num_{c_i}^{b}$
			$\zeta_{c_i}^{b}$
			$num_{c_i}^{s}$
			$\zeta_{c_i}^{s}$
		The quality of the road network	$num_{c_i}^{\tau}$
			v_i
Social interaction	Flow	Inflow	$inflow_{c_i}$
		Outflow	$outflow_{c_i}$
		Intra-flow	$intra-flow_{c_i}$
	Range		λ_i
	Average speed		$\overline{v}_i$

$$\lambda_i = \max_{(O_k, t_{O_k}, D_k, t_{D_k}) \in T^{taxi}} dist(O_k, D_k), \tag{14}$$

where T^{taxi} denotes taxi trajectories and $dist(O_k, D_k)$ denotes the distance between O_k and D_k.

3) Average speed.

The average speed of taxis on roads reflects the fluency of interactions. For a given community c_i, we calculate the average speed $\overline{v}_i$ as

$$\overline{v}_i = \frac{\sum_k \frac{dist(O_k, D_k)}{t_{D_k} - t_{O_k}}}{I\{(O_k, t_{O_k}, D_k, t_{D_k}) \in T^{taxi}\}}, \tag{15}$$

where I denotes the number of trajectories and k is legal when $(O_k, t_{O_k}, D_k, t_{D_k}) \in T^{taxi}$.

Feature Summary

We extract features from BCGD according to the definitions in 4.1 and 4.2. The summary of the extracted features is in **Table 1**. To further capture how the spatial and social features vary over community radius, we set the radius of a community as different distance values (e.g., 0.25, 0.5, 0.75, 1, 1.5, 2, 2.5, and 3 km) and extract a large number of features.

LEARNING TO IDENTIFY THE PATTERNS OF VIBRANT COMMUNITIES

In this section, we present how to select the proper set of important features out of the large number of features obtained from the previous step. We propose a model to spot highly vibrant communities by combining pairwise learning to ranking and sparsity regularization.

Model Description

Since many existing learning-to-rank algorithms use linear rankers, we also learn a linear ranking predictor. Let $\boldsymbol{x_i}$ denote the M-size vector representation of residential community e_i with the above extracted features, f_i denote the predicted vibrancy score, and y_i denote the ground truth of the vibrancy score, then we have

$$f_i(\boldsymbol{x_i}; \boldsymbol{w}) = \sum_{m=1}^{M} w_m x_{im} + \epsilon_i = \boldsymbol{w}^\top \boldsymbol{x_i} + \epsilon_i, \tag{16}$$

where ϵ_i is a zero-mean Gaussian bias with variance σ^2 and $\boldsymbol{w}$ is the weights of the features. In other words, $P(y_i|\boldsymbol{x_i}) = \mathcal{N}(y_i|f_i, \sigma^2) = \mathcal{N}(y_i|\boldsymbol{w}^\top \boldsymbol{x_i}, \sigma^2)$, where $\mathcal{N}$ represents the normal distribution.

Objective Function

While these features indeed capture the spatial configurations and social interactions of residential communities to be ranked, they are often intercorrelated and redundant. These possible confounders lead to poor generalization performance. To address this issue, we adopt a strategy which simultaneously conducts the feature selection while maximizing the ranking accuracy. Since the pairwise ranking strategy is more effective than the listwise ranking strategy, we combine a pairwise ranking objective and a sparsity regularization term in a unified probabilistic modeling framework.

Next, we introduce how to derive the objective for collectively spotting highly vibrant communities and selecting features. Let us denote all parameters by $\Psi = \{\boldsymbol{w}, \boldsymbol{\beta}^2\}$, which are the parameters of the community ranker (we will introduce $\boldsymbol{\beta}^2$ in the following); the hyperparamters by $\Omega = \{a, b, \sigma^2\}$, which are the parameters of sparsity regularization; and the observed data by $\mathcal{D} = \{Y, \Pi\}$, where Y and Π are the community vibrancy scores and ranks of I estates, respectively. For simplicity, we assume the residential communities in $\mathcal{D}$ are sorted and indexed in a descending order of their community vibrancy scores, which compiles a descending ranks as well. In other words, i is both the index and the ranking order of the given community x_i. By Bayesian inference, we have the posterior probability as

$$Pr(\Psi; \mathcal{D}, \Omega) = P(\mathcal{D}|\Psi, \Omega) P(\Psi|\Omega). \tag{17}$$

In **Eq. 17**, the term $P(\mathcal{D}|\Psi, \Omega)$ is the likelihood of the observed data collection $\mathcal{D}$, which can be explained as a joint probability of both community vibrancy scores, $P(Y|\Psi, \Omega)$, and community ranking consistency, $P(\Pi|\Psi, \Omega)$. Here, we treat the ranked list of

communities as a directed graph, $G = <V, E>$, with nodes as communities and edges as pairwise ranking orders. For instance, an edge $i \rightarrow h$ representing community i is ranked higher than community h. From a generative modeling angle, the edge $i \rightarrow h$ is generated by our model through a likelihood function $P(i \rightarrow h)$. The more vibrant the community i is than the community h, the larger $P(i \rightarrow h)$ should be. On the contrary, the case, in which $i \rightarrow h$ but $f_i < f_h$, will punish $P(i \rightarrow h)$. Therefore,

$$\begin{aligned} P(\mathcal{D}|\Psi, \Omega) &= P(Y|\Psi, \Omega)P(\Pi|\Psi, \Omega) \\ &= \prod_{i=1}^{I} \mathcal{N}(y_i|f_i, \sigma^2) \prod_{i=1}^{I-1} \prod_{h=i+1}^{I} P(i \rightarrow h|\Psi, \Omega), \end{aligned} \tag{18}$$

where the generative likelihood of each edge $i \rightarrow h$ is defined as sigmoid $(f_i - f_h)$:

$$P(i \rightarrow h) = \frac{1}{1 + exp(-(f_i - f_h))}. \tag{19}$$

Moreover, the term $P(\Psi|\Omega)$ is the prior of the parameters Ψ. Here, we introduce a sparse weight prior distribution by modifying the commonly used Gaussian prior, such that a different and separate variance parameter β_m^2 is assigned to each weight. Thus, $P(\boldsymbol{w}|\boldsymbol{\alpha}) = \prod_{m=1}^{M} \mathcal{N}(w_m|0, \beta_m^2)$, where β_m^2 represents the variance of the corresponding parameter w_m and $\boldsymbol{\beta}^2 = (\beta_1^2, \ldots, \beta_M^2)^\top$, each of which is treated as a random variable. Later, an inverse gamma prior distribution is further assigned to these hyperparameters, $P(\boldsymbol{\beta}^2|a, b) = \prod_{m=1}^{M} \text{inverse gamma}(\beta_m^2; a, b)$, where a and b are constants and are usually set close to zero. By integrating over the hyperparameters, we obtain a student-t prior for each weight, which is known to enforce sparse representations during learning by setting some feature weights to zero and avoiding overfitting:

$$\begin{aligned} P(\Psi|\Omega) &= P(\boldsymbol{w}|0, \boldsymbol{\beta}^2)P(\boldsymbol{\beta}^2|a, b) \\ &= \prod_{m=1}^{M} \mathcal{N}(w_m|0, \beta_m^2) \prod_{m=1}^{M} Inverse - Gamma(\beta_m^2|a, b). \end{aligned} \tag{20}$$

Parameter Estimation

With the formulated posterior probability, the learning objective is to find the optimal estimation of the parameters Ψ that maximizes the posterior. Hence, by inferring equation 17, we can have the log of the posterior for the proposed model:

$$\begin{aligned} &\mathcal{L}(\boldsymbol{w}, \boldsymbol{\beta}^2|Y, \Pi, a, b, \sigma^2) = \\ &\sum_{i=1}^{I} \left[-\frac{1}{2}\ln\sigma^2 - \frac{(y_i - f_i)^2}{2\sigma^2} \right] + \sum_{i=1}^{I-1} \sum_{h=i+1}^{I} ln \frac{1}{1 + exp(-(f_i - f_h))} \\ &+ \sum_{m=1}^{M} \left[-\frac{1}{2}\ln\beta_m^2 - \frac{w_m^2}{2\beta_m^2} \right] + \sum_{m=1}^{M} \left[-(a+1)\ln\beta_m^2 - \frac{b}{\beta_m^2} \right]. \end{aligned} \tag{21}$$

We apply a gradient descent method to maximize the posterior by updating w_m, β_m^2 through

$$w_m^{(t+1)} = w_m^{(t)} - \epsilon \frac{\partial(-\mathcal{L})}{\partial w_m} \tag{22}$$

and

$$\beta_m^{2(t+1)} = \beta_m^{2(t)} - \epsilon \frac{\partial(-\mathcal{L})}{\partial \beta_m^2}, \tag{23}$$

where

$$\begin{aligned} \frac{\partial(\mathcal{L})}{\partial w_m} &= \sum_{i=1}^{I} \frac{1}{\sigma^2} \left(y_i - \sum_{m=1}^{M} w_m \cdot x_{im} \right) x_{im} \\ &+ \sum_{i=1}^{I-1} \sum_{h=i+1}^{I} \frac{exp(-(f_i - f_h))}{1 + exp(-(f_i - f_h))} (x_{im} - x_{hm}) + \frac{-w_m}{\beta_m^2} \end{aligned} \tag{24}$$

$$\frac{\partial(\mathcal{L})}{\partial \beta_m^2} = \frac{-1}{2\beta_m^2} + \frac{w_m^2}{\beta_m^4} + \frac{-(a+1)}{\beta_m^2} + \frac{b}{\beta_m^4}. \tag{25}$$

EXPERIMENTAL RESULTS

We provide an empirical evaluation of the performances of the proposed method on the real-world residential community–related data.

Data Description

We use the residential community data and crowdsourced geotagged data including bus/subway smart card data, taxi GPS traces data, POIs, and mobile check-in data in Beijing for this study.

Residential Community Data

Since the urban areas of big cities are usually compact due to large population, residential complexes become the major type of properties in the urban area of a city. A residential complex usually includes one or more apartment buildings. We have obtained the data of more than 3,000 Beijing residential complexes by crawling Fang.com, which is the largest real estate online system in China.

Crowdsourced Geotagged Data

- *Taxi GPS Data.* Taxi transits are faster and more expensive and represent an important part of human mobility. Taxi GPS sensors generate trajectory data in the form of sequences of location and time pairs. In our experiments, the taxi GPS traces are collected from a Beijing taxi company from April to August 2012. From the taxi GPS data, we extract the information of each trip, which includes the pick-up location, pick-up time, drop-off location, drop-off time, trip distance, trip speed, driving direction, trip cost, and passenger number.
- *Bus Traces Data.* As two important types of public transit, buses are cheaper with acceptable speeds than taxis that are expensive with faster speed. In urban areas, massive residents choose buses. We have collected Beijing bus trip data through the records of the bus smart card system. Each trip consists of the card id, time stamp,

TABLE 2 | Statistics of the experimental data.

Data source	Properties	Statistics
Taxi GPS	Number of taxis	13,597
	Effective days	92
	Time period	April–August 2012
	Number of trips	8,202,012
	Number of GPS points	111,602
	Total distance (km)	61,269,029
Bus/subway traces	Number of bus/subway stops	9,810
	Time period	August 2012–May 2013
	Number of car holders	300,250
	Number of trips	1,730,000
Mobile check-ins	Number of check-in POIs	5,874
	Number of check-in events	2,762,128
	Number of POI categories	9
	Time period	01/2012-12/2012
POIs	Number of business POIs	328,668
	Positions (longitude and altitude)	328,668
Residential communities	Number of real estates	2,990
	Size of bounding box (km)	40*40

expense, balance, route name, and pick-up and drop-off stop information (names, longitudes, and latitudes).

- *Point of Interest Data.* A point of interest, or POI, is a specific point location that someone may find useful or interesting. We have collected a comprehensive dataset of POI information of Beijing from Dianping and Dajie, including POI name, POI category, latitude, and longitude. The POI categories include catering, shopping, living, sports and leisure, health care, accommodation, scenic spots, business residential, government agencies, science and education, transport facilities, finance and insurance, corporate, and public facilities.
- *Mobile Check-in Data.* Location-based social networks (LBSNs), such as Foursquare, Yelp, and Facebook places, have attracted millions of users to share their digital footprints and opinions with their friends and have enabled us to collect check-ins from mobile apps. Each check-in event typically includes POI name, POI category, address, longitude and latitude, textual comments, and geographic tags. We have collected Beijing check-in data from Weibo, a Chinese version of twitters. It contains $2,762,128$ check-ins in $5,874$ venues.

Table 2 shows the statistics of five data sources.

Baseline Algorithms

To show the effectiveness of our method, we compare our method against the following algorithms.

- RankNet (Burges et al., 2005): It is a combination of a simple probabilistic cost function and using gradient descent methods for learning ranking functions, using a neural network to model the underlying ranking function.
- ListNet (Cao et al., 2007): It is a listwise ranking model with permutation top-k ranking likelihood as objective function. ListNet introduces two probability models, respectively, referred to as permutation probability and top-k probability, to define a listwise loss function for learning. Neural network and gradient descent are then employed as model and algorithm in the learning method.
- Coordinate Ascent (Dang and Croft, 2010): It uses a loss function called the domination loss. Coordinate ascent extends the loss by incorporating margin requirements over pairs of instances and enables the usage of multivalued feedback. Coordinate ascent devises a simple yet effective coordinate descent algorithm that is guaranteed to converge to the unique optimal solution.
- Random Forests (Jiang, 2011): It is a ranking strategy through learning the predictions from an ensemble of random trees.

In the experiments, we utilize RTree[1] to index geographic items (i.e., taxi and bus trajectories) and extract the defined features. We use Jieba,[2] which is a Chinese/English text segmentation module to segment words and extract tags.

For traditional LTR algorithms, we use RankLib.[3] We set the number of training epochs to 100, the number of hidden layers to 1, the number of hidden nodes per layer to 10, and the learning rate to 0.00005 for RankNet. We set the number of iterations to 300 and the number of threshold candidates to 10 for RankBoost. We set number of random restarts to 5, the number of iterations to search in each dimension to 25, and tolerance to 0.001 for Coordinate Ascent. We set the number of bags to 300, the number of leaves to 10, the number of threshold candidates to 256, the number of leaves for each tree to 100, and the learning rate to 0.1 for Random Forest. We set a to 0.001, b to 0.001, and σ^2 to 1,000 for our model.

All the codes are implemented in Python, including modeling, feature extraction, and visualization. All codes can be

[1]https://pypi.python.org/pypi/Rtree/

[2]https://github.com/fxsjy/jieba

[3]http://sourceforge.net/p/lemur/wiki/RankLib/

downloaded *via* the link.[4] And all the evaluations are performed on a x64 machine with i7 2.50 GHz Intel CPU (with four cores) and 16 GB RAM. The operation system is OS X EI Capitan.

Evaluation Metrics

To evaluate the effectiveness of the proposed model, we use the following metrics.

- Normalized Discounted Cumulative Gain (NDCG@N).

The discounted cumulative gain (DCG@N) is given by

$$DCG[n] = \begin{cases} rel_n & if\ n = 1 \\ DCG[n-1] + \frac{rel_n}{log_2 n}, & if\ n >= 2 \end{cases}, \quad (26)$$

where rel_n denotes the vibrancy grade level of the n-th community, defined in **Eq. 2**. Later, given the ideal discounted cumulative gain DCG', NDCG at the n-th position can be computed as

$$NDCG[n] = \frac{DCG[n]}{DCG'[n]}. \quad (27)$$

The larger NDCG@N is, the higher the top-N ranking accuracy the classifier has.

- Kendall's Tau coefficient.

Kendall's Tau coefficient (or Tau for short) measures the overall ranking accuracy. Let us assume that each community i is associated with a benchmark vibrancy y_i and a predicted vibrancy score f_i. Then, for a community pair $< i, j >$, $< i, j >$ is said to be concordant, if both $y_i > y_j$ and $f_i > f_j$ or if both $y_i < y_j$ and $f_i < f_j$. Also, $< i, j >$ is said to be discordant, if both $y_i < y_j$ and $f_i > f_j$ or if both $y_i < y_j$ and $f_i > f_j$. Tau is given by

$$\text{Tau} = \frac{\#_{conc} - \#_{disc}}{\#_{conc} + \#_{disc}}, \quad (28)$$

where $\#_{conc}$ denotes the concordant pairs and $\#_{disc}$ denotes the discordant pairs.

- Recall.

Since we use a six-level rating system ($5 > 4 > 3 > 2 > 1 > 0$) instead of the binary rating, we treat the rating ≥ 5 as "highly vibrant" and the rating < 5 as "fairly vibrant." Given a top-N estate list E_N sorted in a descending order of the prediction values, the recall is defined as

$$\text{Recall@}N = \frac{|E_N \cap E_{\geq 5}|}{|E_{\geq 5}|}, \quad (29)$$

where $E_{\geq 5}$ are the estates whose ratings are greater or equal to 5.

[4]https://www.dropbox.com/s/tyamms9625aivtk/code.py?dl=0

Analysis of Scoring Community Vibrancy

We calculate the vibrancy scores of residential communities in the dataset based on the proposed metric **Eq. 2**. After that, all the communities are sorted in a descending order in terms of vibrancy scores, as shown in **Figure 3A**. We can observe that there are some fault ages on the curve, where the vibrancy scores of some communities significantly increase, whereas the vibrancy scores of many communities remain stable. To prepare the grade levels of community vibrancy for our ranking framework, we utilize these inflection points. First, we identify five inflection points in the curve, which, respectively, denote the vibrancy scores of 0.7713, 0.4685, 0.3375, 0.1506, 0.0523, and 0.7713. The five inflection points split the curve into six segments. After that, we assign six-level ratings to each segment as its ranking relevance label, for instance, 5, 4, 3, 2, 1, and 0, respectively, in a descending order based on the vibrancy scores. As a result, we obtain six rating levels for the ranking process, as shown in **Figures 3B,C**.

The curve in **Figure 3A** shows that the distribution of the community vibrancy scores complies with a power law distribution, indicating only a small number of residential communities are highly vibrant, and most communities are around the mean value of the vibrancy scores. This observation is consistent with our common sense about our world: Most people are middle class and only a small number are rich. The six rating levels are shown in **Figures 3B,C**, which visualizes the distribution of the six vibrancy levels of all the communities.

Correlation Analysis of Features

We provide a visualization analysis to validate the correlation between the extracted features and the vibrancy scores of communities. We use the scatter plot matrix for correlation analysis. Each non-diagonal chart in a scatter plot matrix shows the correlation between a pair of features whose feature names are listed in the corresponding diagonal charts. Given a set of N features, there are N-choose-two pairs of features, and thus the same numbers of scatter plots. The dots represent the communities and their colors represent the levels of vibrancy values. For readability, we use $R6 > R5 > R4 > R3 > R2 > R1$ (symbol) to represent $5 > 4 > 3 > 2 > 1 > 0$ (number) in **Figure 4**. For detailed quantitative results, refer to **Table 3**.

In **Figure 4A**, we present the correlation between bus trace features (inflow interaction, outflow interaction, intra-flow interaction, distance to bus stops, and the density of) and vibrancy values of communities. As can be seen, the R5 communities tend to appear at the top right corner of all the non-diagonal charts. However, the R6 communities appear at the middle of the figure. This implies that the bus is the major transportation for common communities, while people tend to visit top vibrant and high-end communities by other kinds of vehicles.

In **Figure 4B**, we show the positive correlation between the taxi inflow, outflow, and intra-flow volumes of communities and vibrancy values. This shows that the taxi is an important transportation to visit vibrant communities, which is consistent with the observation of buses in **Figure 4A**.

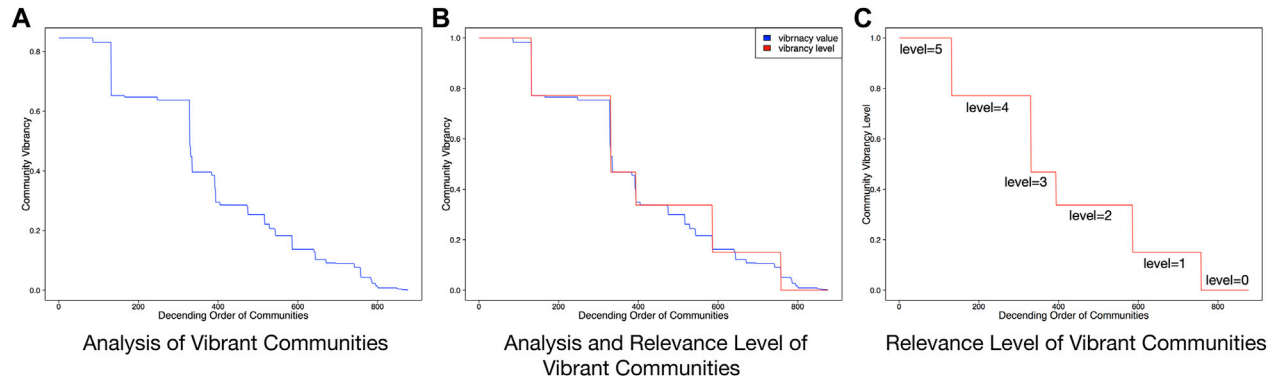

FIGURE 3 | Analysis of urban vibrancy based on the proposed metric.

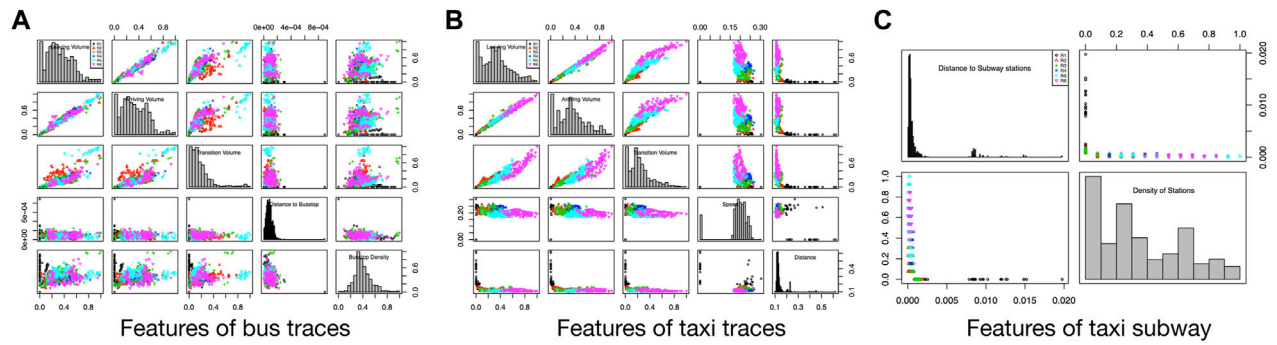

FIGURE 4 | Feature correlation analysis of bus traces, taxi traces, and subways.

TABLE 3 | Feature correlation analysis of bus traces, taxi traces, and subways.

	R6	R5	R4	R3	R2	R1
Inflow of bus	0.63	0.85	0.36	0.11	0.56	0.17
Outflow of bus	0.59	0.88	0.34	0.57	0.49	0.24
Intra-flow of bus	0.67	0.92	0.31	0.37	0.28	0.05
Distance to bus stop	0.21	0.57	0.86	0.74	0.42	0.45
Bus stop density	0.14	0.09	0.88	0.18	0.50	0.65
Inflow of taxi	0.89	0.73	0.79	0.59	0.14	0.51
Outflow of taxi	0.94	0.16	0.41	0.12	0.61	0.05
Intra-flow of taxi	0.85	0.62	0.84	0.49	0.42	0.27
Speed of taxi	0.92	0.13	0.26	0.01	0.65	0.88
Traveling distance of taxi	0.89	0.36	0.51	0.33	0.80	0.25
Distance to subway station	0.89	0.63	0.47	0.10	0.07	0.01
Subway station density	0.93	0.82	0.45	0.19	0.08	0.02

However, the commute distances of taxis have a negative correlation with the vibrancy scores. In other words, the shorter the commute distances of taxis are, the higher the vibrancy scores of residential communities are. A potential interpretation of this observation is that since taxis are valued by white-collar and business people, the destinations of taxi trajectories usually are important places (i.e., conference centers, business hotels, companies, and government organizations). If the commute distance of taxis is shorter, the targeted neighborhood is closer to these important places.

In **Figure 4C**, we show the power law correlation between the community vibrancy scores and the subway-related features, including the distance to the subway stations and the density of the subway stations. We can obtain the observation similar to **Figure 4A** that subways are not the most important transportation for visiting top vibrant communities. Based on the observations in **Figures 4A,B**, we can find that the public transportation (i.e., bus and subway) has huge effects on the communities of R4 and R5. However, the influence of the public transportation on top vibrant communities is small. The taxi-related features show nearly a positive linear relation with the community vibrancy scores, especially for top vibrant communities (R6). There may be an explanation that if a community is very vibrant, the cost spent on transportation is likely to be high. As known to all, public transportations are relatively slow but cheap. Taxis are expensive but fast. Therefore, the high-consumption group (like white-collar and business people) who can afford taxis are more in favor of taxis.

In summary, the visualization results show the correctness of our intuitions about defining and extracting discriminative features.

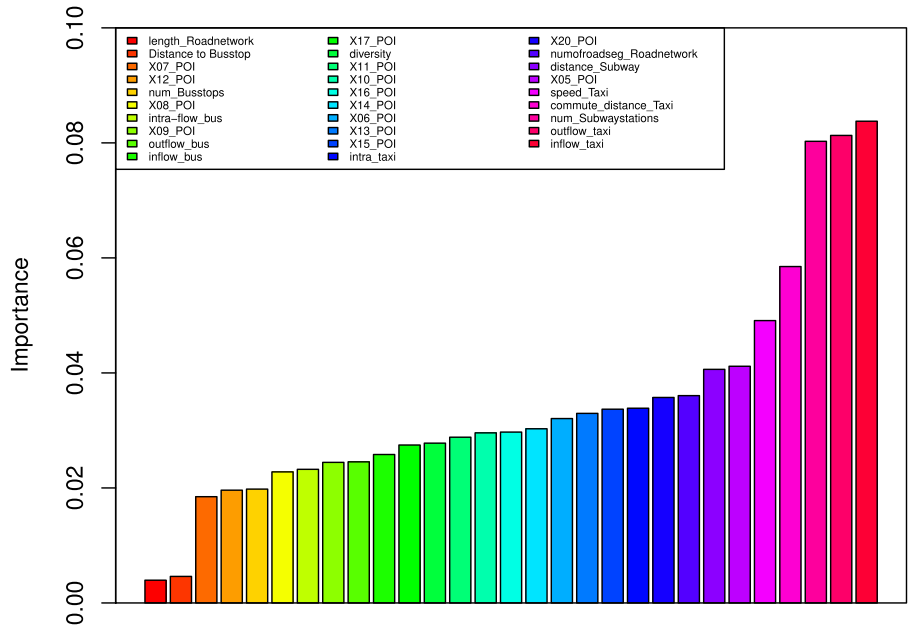

FIGURE 5 | Feature importances based on information gain.

Examining the Importance of Urban Geography and Human Mobility Features

We measure the information gain of each feature described in the section *Discovering Patterns of Vibrant Communities* to understand the importance of the spatial and mobility patterns in community vibrancy. Specifically, we calculate the information gain of each feature for each vibrancy levels (i.e., $5 > 4 > 3 > 2 > 1 > 0$) across our data. **Figure 5** shows the results of the information gain analysis for a decision tree classifier.

We have some interesting observations from **Figure 5**:

- Taxi-related features including inflow, outflow, commute distance, and speed are top ranked as 0.084, 0.081, 0.058, and 0.049, respectively. Surprisingly, the intra-flow of taxi is ranked as 0.034 in the middle of the list. This conforms with our common sense that the human mobility across communities encodes both specific trip purposes and the destinations that can meet people's demands. That is exactly why the vibrant communities can attract people. However, the mobility within communities cannot show the sign explicitly.
- The information gain of POI-related features distributes in the range of the list. The highest is 0.041, while the lowest is 0.018. The reason for such big differences can be that different POI categories always have different functionalities. Some POIs, like shopping and restaurants, are popular to people and can provide the recreation and entertainment functionality, while some POIs, like vehicle services, would not appear too many in our daily life. Therefore, specific POI categories may contribute a lot to the community vibrancy but some may not.
- The public transportation–related features including the distance to bus stops, the number of bus stops, the intra-flow of buses, the outflow of buses, the distance to subway stations, and the number of subway stations are ranked at 0.005, 0.020, 0.023, 0.025, 0.034, 0.041, and 0.080, respectively. Moreover, the subway-related features are more important than the bus-related features. There is a possible explanation that the subway is much more rapid than the bus and we also do not need to worry about the traffic jam on the subway. In this case, the more rapid and convenient subway outweighs.
- For road network–related features, that is, the length of road networks and the number of intersections, the information gain value is 0.004 and 0.036. We need to notice that the information gain of the number of intersections nearly catch up with taxi-related features. This is because more intersections mean that it is more likely to take a taxi. Convenient transportation facilities in the vibrant communities always attract many people to visit.

Model Performance Comparison

We compare the performance of our method with four baseline algorithms in terms of Tau and NDCG.

In **Table 4**, we list details of performance of different models. Our method achieves 0.6081 NDCG@3, 0.5283 NDCG@5, 0.3736 NDCG@10, and 0.3314 NDCG@15, which obviously outperforms the baseline algorithms with a significant margin. Our method fuses sparsity regularization and pairwise ranking objective and offers an increase in comparison to RankNet which has the best performance in baseline algorithms, as shown in **Figure 6**.

TABLE 4 | Performance comparison of our approach and baselines.

	Random Forests	ListNet	Coordinate Ascent	RankNet	Our model
NDCG@3	0.0867	0.1002	0.0788	0.2	0.7103
NDCG@5	0.0879	0.0997	0.0841	0.2	0.5897
NDCG@10	0.0919	0.1003	0.0861	0.2	0.4544
NDCG@15	0.0907	0.1004	0.0852	0.2	0.3908
Tau	−1.0	−0.0401	−0.0616	−0.4699	−0.4594

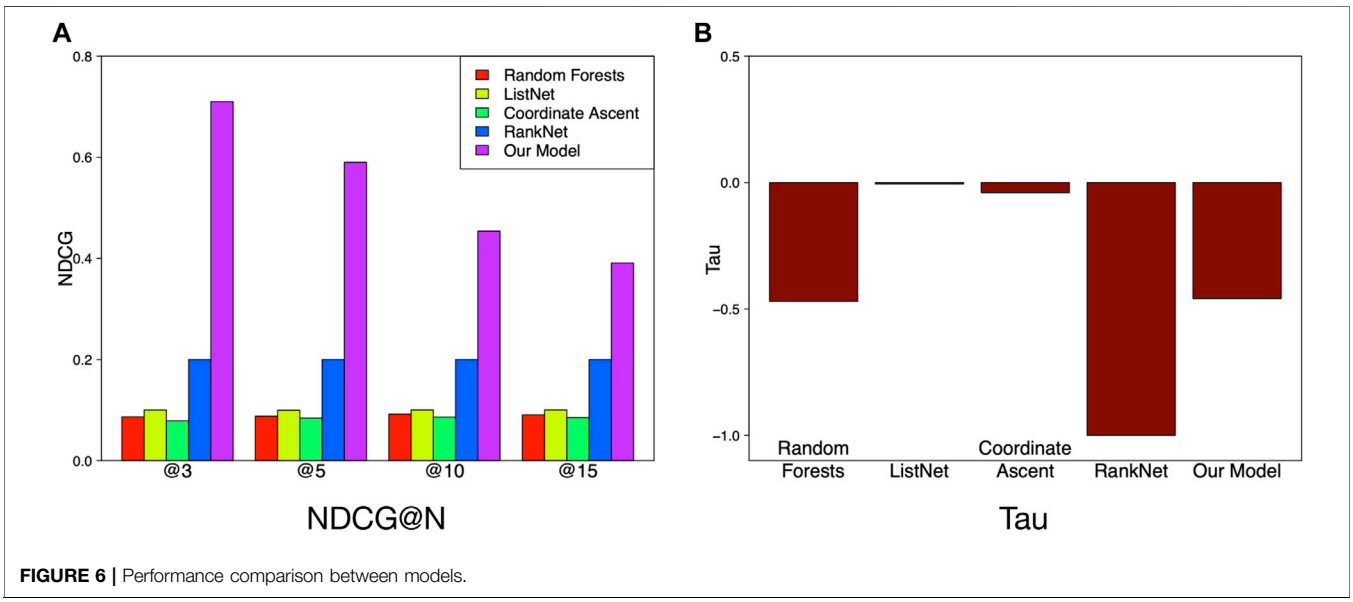

FIGURE 6 | Performance comparison between models.

This observation validates the superiority of our method when considering many intercorrelative features with confounders. Moreover, the effectiveness of considering both sparsity regularization and ranking accuracy is proved.

With respect to the overall ranking, our method achieves the highest Tau (0.6137). Surprisingly, all the baselines perform badly on Tau where values of Tau are all negative. The observation indicates that the number of concordant pairs is slightly less than the number of discordant pairs, which demonstrates the lower accuracies of the baseline algorithms on the whole ranking list. However, our method achieves a balanced performance in both top-k and overall ranking.

Another fact we draw from **Figure 6** is that the NDCG of our model increases with N getting small, which indicates that the ranking performance of our model does well in the top-k ranking task, especially for the very top part.

Feature Performance Comparison

We evaluate the performances of different features segmented from two angles. The feature performance is evaluated in terms of NDCG@N, Recall@N, and Tau, respectively.

- Evaluation on features of different categories.

At the beginning of the article, we emphasize that the vibrancy of community is valued in terms of spatial configurations and social interactions. The difference between spatial configurations and social interactions is that spatial configuration represents the static state of a community, implying the geographical representations and distributions of static geographic items, like POIs and bus stops, whereas social interactions represent the dynamics of a community, showing the mobility patterns of mobile objects, like taxis and buses. Therefore, we split features into these two categories. As shown in **Figure 7A**, compared to spatial configuration features, social interaction features perform better. This observation shows that dynamic features contribute more to the ranking accuracy of our model. It is very necessary to study the social interaction features to further explore more useful dynamic patterns for improving ranking performances. Besides, **Figures 7B,C** also provide other evidences to validate the better performance of the social interaction features compared with the spatial configuration features in terms of Recall@N and Tau.

- Evaluation on features of different data sources.

Aside from studying the categories of features, we also study the performances of different data sources as we have collected data from taxi GPS trajectories, bus GPS trajectories, road networks, and POIs. Here, we segment the extracted features in terms of different data sources and investigate which source is more effective for ranking urban vibrancy. **Figure 8A** shows the taxi and POI–related features contribute most to the accuracy of

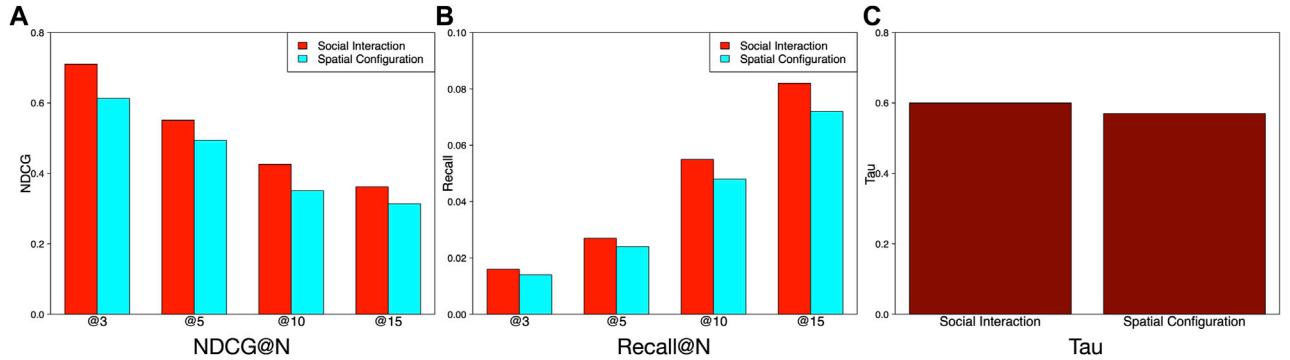

FIGURE 7 | Comparison of feature performance based on different states.

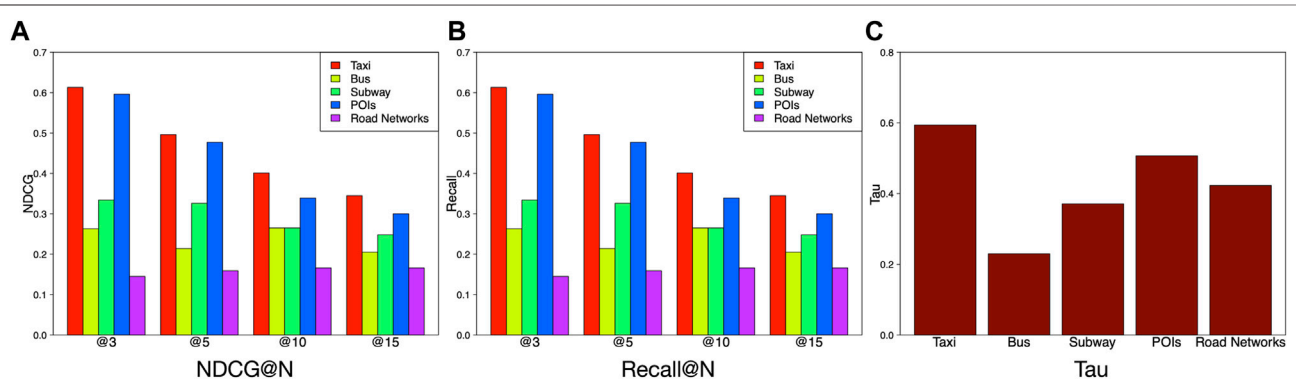

FIGURE 8 | Comparison of feature performance based on data sources.

the proposed model, while the road network–related features contribute the least to the model accuracy. Moreover, taxi data and POI data are the two major sources to represent the social interactions and spatial configurations, respectively. This observation is consistent with the result in **Figures 7**, **8B** which show taxis data performance is the best among all the data sources. The following are bus, subway, POIs, and road networks, respectively. As for Tau, the performances of different data sources are ranked in a descending order as taxi > POIs > road networks > subway > bus. Overall, taxi-related data are the most useful source to construct effective features. Besides, when we examine the performances of all kinds of transportation-related data, taxi > subway > bus. Here is a possible explanation. Bus is the most common type for commuting. Also, bus and subway are for massive people traveling in a certain planed trajectories based on schedule. However, taxis are for personal usage, making the range of traditional zone unlimited. Therefore, we can dig more information from taxi-related features.

- Evaluation on features of different radius distances.

We segment the features in terms of different radius of communities and investigate the proper radius of neighborhoods for ranking community vibrancy. **Figure 9** shows the performance comparisons of the feature sets of different radius distances (i.e., 0.25, 0.5, 0.75, 1, 1.5, 2, and 3 km). We observe that the radius distance of neighborhood can affect the ranking performance. **Figure 9A** shows that the NDCGs for 0.5, 0.75, 1, 1.5, 2.0, and 3 km are almost same, while the radius of 0.25 km shows a slightly higher NDCG. However, the high NDCGs of 0.5–3 km are used to consistently validate the superiority of our model. For the recall performances in **Figure 9B**, we can obtain an interesting observation that there is a descending trend when the radius is getting larger. This may be due to the fact that more data are available when the community radius is larger. Abundant data result in poor generalization of a model and lead to the descending trend of the ranking accuracy. **Figure 9C** implies that the Tau values vary slightly in the interval of $[-0.3, -0.6]$ when the radius of communities drops $(0.25km, 0.5km, 0.75km, 1km, 1.5km, 2.0km, 3km)$. This reflects the robustness of our method from another perspective.

Based on the above analysis, we should not set the radius of communities too small (i.e., $0.25km$) because of the information limitation. On the other hand, too large value of radius is useless due to the stability of the ranking performance. Therefore, we set the radius as $1\ km$ in this study.

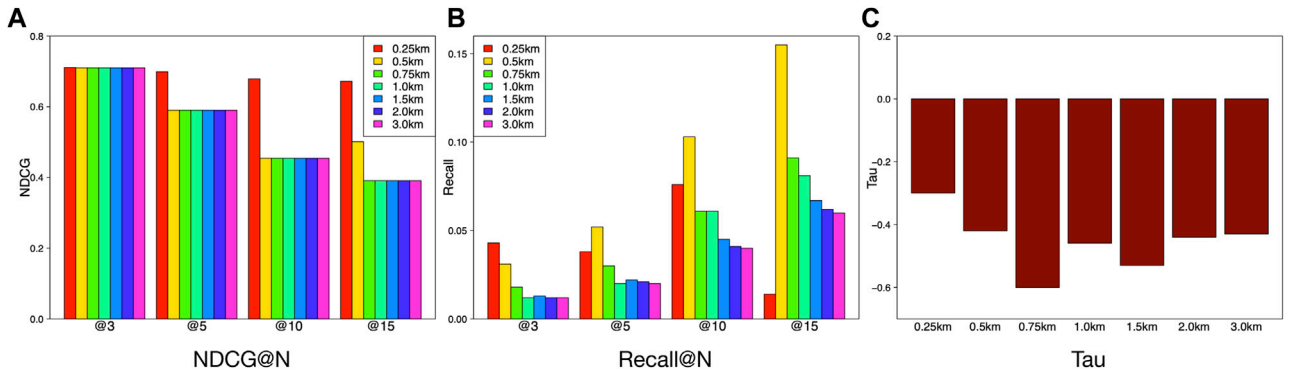

FIGURE 9 | Comparison of feature performance based on different radius distance.

RELATED WORK

Urban Planning

Researchers have developed conceptual and empirical measurements on urban vibrancy from different aspects. The first aspect is density. The work in Glaeser et al. (2001) pointed out modern cities will be consumer-centric rather than production-centric; the future of cities depends on the demand for urban *density*. Couture *et al.* found that high-density areas benefit residents in terms of more social interaction and diverse consumption opportunities, and people are willing to pay higher rents and transportation costs for high-density places (Couture, 2013). The second aspect is diversity. Farber *et al.* found that proper urban structure leads to spatial concentration of residents and diversity of products and services (Farber and Li, 2013). Talen *et al.* found that mixed land uses can encourage workability and foster social interaction (Talen, 1999). Malizia *et al.* found that vibrant communities are usually compact, dense, and accessible with diverse land uses (Malizia and Song, 2014). Neutens *et al.* found that high-density and mixed land uses can benefit quality social interaction and enhance vibrancy (Neutens et al., 2013). The third aspect is human-related dynamic factors. Dougal *et al.* argued urban vibrancy should be measured by dynamic human-dependent factors that vary over time (Dougal et al., 2015). For example, Farber and Li (2013) proposed social interaction potential as a measurement; Audretsch et al. (2003) proposed the knowledge diffusion among workers as a metric; Jaffe et al. (1993) measured vibrancy with technology spillovers between neighboring firms; Glaeser et al. (2001) used consumption externalities between its residents as a metric; and the work by Dougal et al. (2015) devised firm investment opportunities as a metric. In summary, prior studies found that 1) urban vibrancy is nearly always related to density and diversity in terms of both static geographical and dynamic human-related factors; and 2) urban vibrancy is complex and should include density, diversity, and human activities.

Urban Computing With Geography and Mobility

Urban computing (Zheng et al., 2014) is a process of acquisition, integration, and analysis of urban data (e.g., sensors, devices, vehicles, buildings, and human) to tackle the major issues that cities face. Our work also has a connection with mining mobile, geography, and mobility data to tackle issues in urban space. Tseng et al. mine the behavior patterns from mobile sensor data to enhance system performance (Tseng and Lin, 2006). The work by Ceci et al. (2007) identifies emerging patterns with multirelational approach from spatial data. Liu et al. detect spatiotemporal causality of outliers in traffic data (Liu et al., 2011). Yuan et al. discover regional functions of a city using POIs and taxi traces (Yuan et al., 2012). Heierman et al. mine the device usage patterns of homeowners for smart houses (Heierman and Cook, 2003). The study by Karamshuk et al. (2013) selects the optimal sites for retail stores by mining Foursquare data. Zheng et al. (2014) mine the driving route for end users by considering the physical feature of a route, traffic flow, and driving behavior.

Learning-to-Rank

Our work can be categorized into learning-to-rank (LTR), which includes pointwise, pairwise, and listwise approaches (Li, 2011). The pointwise methods (Li, 2011) reduce the LTR task to a regression problem: given a single query–document pair, it predicts its score. The pairwise methods reduce the LTR task to a classification problem. The goal of the pairwise ranking is to learn a binary classifier to identify the better document in a given document pair by minimizing the average number of inversions in ranking, for example, RankNet (Burges et al., 2005), RankBoost (Freund et al., 2003), RankSVM (Herbrich et al., 2000), and LambdaRank (Burges et al., 2007). The listwise methods optimize a ranking loss metric over lists instead of document pairs (Xia et al., 2008). For instance, H. Li et al. propose AdaRank (Xu and Li, 2007) and ListNet (Cao et al., 2007) and Burges et al. propose LambdaMART (Burges, 2010). The recent work by Agarwal et al. (2012) and Agrawal et al. (2006) further studied multifaceted ranking and context-sensitive ranking. The work by Rendle et al. (2009), Weng and Lin (2011), and Gantner et al. (2012) provide full Bayesian explanations and optimize the posterior of pointwise, pairwise, and listwise ranking models, respectively. The study by Shi et al. (2013) unifies both rating error and ranking error as objective function to enhance top-k recommendation. More recent work (Lai et al., 2013) further

learns the ranking model which is constrained to be with only a few nonzero coefficients using L1 constraint and proposes a learning algorithm from the primal dual perspective.

CONCLUSION REMARKS

In this article, we aimed to measure urban vibrancy by examining spatial configuration and social interaction of communities with Big Crowdsourced Geotagged Data. We proposed a fused scoring framework, combining diversity and density of consumer activities with F-1 score. We extracted features to represent spatial configuration and social interaction, respectively. To learn vibrancy values based on the proposed scoring framework, we designed a sparse ranking model which is mutually enhanced by simultaneously conducting feature selection and maximizing communities' vibrancy ranking accuracy. Finally, the experimental results with BCGD demonstrate the competitive effectiveness of both extracted features and learning models. With the high accuracy ranking prediction, we explore the potential to use BCGD for providing useful strategies for governments on urban planning. On the other hand, higher vibrancy leads to more consumers and the high quantity of consumers enhance vibrant communities, which invents a virtuous cycle for the development of cities.

DATA AVAILABILITY STATEMENT

The data analyzed in this study are subject to the following licenses/restrictions: Data belong to Microsoft. Requests to access these datasets should be directed to yanjie.fu@ucf.edu.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This research was supported by the National Science Foundation (NSF) via the Grant Numbers: 2045567, 2006889, 2040950, and 1947534.

REFERENCES

Agarwal, D., Chen, B.-C., and Wang, X. (2012). "Multi-faceted Ranking of News Articles Using post-read Actions," in Proceedings of the 21st ACM international conference on Information and knowledge management, Maui, Hawaii, October 29–November 2, 2012 (ACM), 694–703.

Agrawal, R., Rantzau, R., and Terzi, E. (2006). "Context-sensitive Ranking," in Proceedings of the 2006 ACM SIGMOD international conference on Management of data, Chicago, IL, June 27–29, 2006 (New York, NY: ACM), 383–394.

Audretsch, D. B., Feldman, M., Henderson, J. V., and Thisse, J.-F. (2003). *Handbook of Urban and Regional Economics*. Amsterdam, Netherlands: Elsevier. 4.

Burges, C. J. (2010). From Ranknet to Lambdarank to Lambdamart: An Overview. *Learning* 11, 81.

Burges, C. J., Ragno, R., and Le, Q. V. (2007). Learning to Rank with Nonsmooth Cost Functions. *Adv. Neural Inf. Process. Syst.*, 193–200.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., et al. (2005). "Learning to Rank Using Gradient Descent," in Proceedings of the 22nd international conference on Machine learning, Bonn, Germany, August 7–11, 2005 (New York, NY: ACM), 89–96.

Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., and Li, H. (2007). "Learning to Rank: from Pairwise Approach to Listwise Approach," in Proceedings of the 24th international conference on Machine learning, Corvalis, OR, June 20–24, 2007 (New York, NY: ACM), 129–136.

Ceci, M., Appice, A., and Malerba, D. (2007). "Discovering Emerging Patterns in Spatial Databases: A Multi-Relational Approach," in European Conference on Principles of Data Mining and Knowledge Discovery, Warsaw, Poland, September 17–21, 2007 (Springer), 390–397.

Church, R. L., and Murray, A. T. (2009). *Business Site Selection, Location Analysis, and GIS*. Wiley Online Library.

Collins, P., Al-Nakeeb, Y., Nevill, A., and Lyons, M. (2012). The Impact of the Built Environment on Young People's Physical Activity Patterns: A Suburban-Rural Comparison Using GPS. *Ijerph* 9, 3030–3050. doi:10.3390/ijerph9093030

Couture, V. (2013). *Valuing the Consumption Benefits of Urban Density*. Berkeley: University of California. Processed. doi:10.4324/9780203057513

Dang, V., and Croft, B. (2010). "Feature Selection for Document Ranking Using Best First Search and Coordinate Ascent," in Sigir workshop on feature generation and selection for information retrieval, Geneva, Switzerland, July 19–23, 2010. doi:10.1145/1718487.1718493

Dougal, C., Parsons, C. A., and Titman, S. (2015). Urban Vibrancy and Corporate Growth. *J. Finance* 70, 163–210. doi:10.1111/jofi.12215

Evans, G., Foord, J., Porta, S., Thwaites, K., Romice, O., and Greaves, M. (2007). "The Generation of Diversity: Mixed-Use and Urban Sustainability," in *Urban Sustainability through Environmental Design: Approaches to Time People-Place Responsive Urban Spaces*, 95–101.

Farber, S., and Li, X. (2013). Urban Sprawl and Social Interaction Potential: an Empirical Analysis of Large Metropolitan Regions in the united states. *J. Transport Geogr.* 31, 267–277. doi:10.1016/j.jtrangeo.2013.03.002

Farber, S., Neutens, T., Carrasco, J.-A., and Rojas, C. (2014). Social Interaction Potential and the Spatial Distribution of Face-To-Face Social Interactions. *Environ. Plann. B Plann. Des.* 41, 960–976. doi:10.1068/b120034p

Farber, S., Neutens, T., Miller, H. J., and Li, X. (2013). The Social Interaction Potential of Metropolitan Regions: A Time-Geographic Measurement Approach Using Joint Accessibility. *Ann. Assoc. Am. Geogr.* 103, 483–504. doi:10.1080/00045608.2012.689238

Freund, Y., Iyer, R., Schapire, R. E., and Singer, Y. (2003). An Efficient Boosting Algorithm for Combining Preferences. *J. machine Learn. Res.* 4, 933–969.

Gantner, Z., Drumond, L., Freudenthaler, C., and Schmidt-Thieme, L. (2012). "Personalized Ranking for Non-uniformly Sampled Items," in Proceedings of KDD Cup 2011, San Diego, CA, August 21–24, 2011, 231–247.

Glaeser, E. L., Kolko, J., and Saiz, A. (2001). Consumer City. *J. Econ. Geogr.* 1, 27–50. doi:10.1093/jeg/1.1.27

Heierman, E. O., and Cook, D. J. (2003). "Improving home Automation by Discovering Regularly Occurring Device Usage Patterns," in Data Mining, 2003. ICDM 2003. Third IEEE International Conference on, Melbourne, FL, December 19-22, 2003 (IEEE), 537–540.

Herbrich, R., Graepel, T., and Obermayer, K. (2000). "Large Margin Rank Boundaries for Ordinal Regression," in *Advances in Large Margin Classifiers*. Editors A. J. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans (Cambrodge, United Kingdom: MIT Press).

Jaffe, A. B., Trajtenberg, M., and Henderson, R. (1993). Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations. *Q. J. Econ.* 108, 577–598. doi:10.2307/2118401

Jiang, L. (2011). Learning Random Forests for Ranking. *Front. Comput. Sci. China* 5, 79–86. doi:10.1007/s11704-010-0388-5

Karamshuk, D., Noulas, A., Scellato, S., Nicosia, V., and Mascolo, C. (2013). "Geo-spotting: Mining Online Location-Based Services for Optimal Retail Store Placement," in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, Chicago, IL, August 11–14, 2013 (New York, NY: ACM), 793–801.

Khan, M., M. Kockelman, K., and Xiong, X. (2014). Models for Anticipating Non-motorized Travel Choices, and the Role of the Built Environment. *Transport Policy* 35, 117–126. doi:10.1016/j.tranpol.2014.05.008

Koster, H. R. A., and Rouwendal, J. (2012). The Impact of Mixed Land Use on Residential Property Values*. *J. Reg. Sci.* 52, 733–761. doi:10.1111/j.1467-9787.2012.00776.x

Lai, H., Pan, Y., Liu, C., Lin, L., and Wu, J. (2013). Sparse Learning-To-Rank via an Efficient Primal-Dual Algorithm. *IEEE Trans. Comput.* 62, 1221–1233. doi:10.1109/tc.2012.62

Li, H. (2011). A Short Introduction to Learning to Rank. *IEICE Trans. Inf. Syst.* E94-D, 1854–1862. doi:10.1587/transinf.e94.d.1854

Liu, W., Zheng, Y., Chawla, S., Yuan, J., and Xing, X. (2011). "Discovering Spatio-Temporal Causal Interactions in Traffic Data Streams," in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, San Diego, CA, August 21-24, 2011 (New York, NY: ACM), 1010–1018.

Loehr, S. (2013). Mixed-use, Mixed Impact: Re-examining the Relationship between Non-residential Land Uses and Residential Property Values. Master Thesis. New York (NY): Columbia University.

Malizia, E., and Song, Y. (2014). Vibrant Downtowns: Can Vibrancy Explain Variations in Downtown Property Performance?

Neutens, T., Farber, S., Delafontaine, M., and Boussauw, K. (2013). Spatial Variation in the Potential for Social Interaction: A Case Study in flanders (belgium). *Comput. Environ. Urban Syst.* 41, 318–331. doi:10.1016/j.compenvurbsys.2012.06.007

Rendle, S., Freudenthaler, C., Gantner, Z., and Schmidt-Thieme, L. (2009). "Bpr: Bayesian Personalized Ranking from Implicit Feedback," in Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence, Montreal, QC, June 18–21, 2009 (New York, NY: AUAI Press), 452–461.

Shi, Y., Larson, M., and Hanjalic, A. (2013). Unifying Rating-Oriented and Ranking-Oriented Collaborative Filtering for Improved Recommendation. *Inf. Sci.* 229, 29–39. doi:10.1016/j.ins.2012.12.002

Song, Y., and Knaap, G.-J. (2004). Measuring the Effects of Mixed Land Uses on Housing Values. *Reg. Sci. Urban Econ.* 34, 663–680. doi:10.1016/j.regsciurbeco.2004.02.003

Talen, E. (1999). Sense of Community and Neighbourhood Form: An Assessment of the Social Doctrine of New Urbanism. *Urban Stud.* 36, 1361–1379. doi:10.1080/0042098993033

Tseng, V. S., and Lin, K. W. (2006). Efficient Mining and Prediction of User Behavior Patterns in mobile Web Systems. *Inf. Softw. Technol.* 48, 357–369. doi:10.1016/j.infsof.2005.12.014

Weng, R. C., and Lin, C.-J. (2011). A Bayesian Approximation Method for Online Ranking. *J. Machine Learn. Res.* 12, 267–300.

Xia, F., Liu, T.-Y., Wang, J., Zhang, W., and Li, H. (2008). "Listwise Approach to Learning to Rank: Theory and Algorithm," in Proceedings of the 25th international conference on Machine learning, Helsinki, Finland, July 5–9, 2008 (New York, NY: ACM), 1192–1199.

Xu, J., and Li, H. (2007). "Adarank: a Boosting Algorithm for Information Retrieval," in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, Amsterdam, Netherlands, July 23-27, 2007 (New York, NY: ACM), 391–398.

Yuan, J., Zheng, Y., and Xie, X. (2012). "Discovering Regions of Different Functions in a City Using Human Mobility and Pois," in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, Beijing, China, August 12–16, 2012 (New York, NY: ACM), 186–194.

Yue, Y., Zhuang, Y., Yeh, A. G. O., Xie, J.-Y., Ma, C.-L., and Li, Q.-Q. (2017). Measurements of Poi-Based Mixed Use and Their Relationships with Neighbourhood Vibrancy. *Int. J. Geographical Inf. Sci.* 31, 658–675. doi:10.1080/13658816.2016.1220561

Zhang, M. (2005). Exploring the Relationship between Urban Form and Nonwork Travel through Time Use Analysis. *Landscape Urban Plann.* 73, 244–261. doi:10.1016/j.landurbplan.2004.11.008

Zheng, Y., Capra, L., Wolfson, O., and Yang, H. (2014). Urban Computing: Concepts, Methodologies, and Applications. *ACM Trans. Intell. Syst. Technol. (Tist)* 5, 38. doi:10.1145/2629592

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.