Participatory Cultural Mapping Based on Collective Behavior Data in Location Based Social Networks

DINGQI YANG, eXascale Infolab, University of Fribourg
DAQING ZHANG, Institut Mines-Télécom/Télécom SudParis and Peking University
BINGQING QU, University of Rennes 1 - IRISA & INRIA Rennes

Culture has been recognized as a driving impetus for human development. It co-evolves with both human belief and behavior. When studying culture, *Cultural Mapping* is a crucial tool to visualize cultures in different aspects (e.g., religions and languages) on the map from the perspectives of indigenous and local people. The existing cultural mapping approaches usually rely on large-scale survey data with respect to human belief, such as moral values. However, such a data collection method not only incurs a significant cost of both human resources and time, but also fails to capture human behavior, which massively reflects cultural information. In addition, it is practically difficult to collect large-scale human behavior data. Fortunately, with the recent boom of Location Based Social Networks (LBSNs), a considerable number of users report their activities in LBSNs in a participatory manner, which provides us with an unprecedented opportunity to study large-scale user behavioral data. In this paper, we propose a participatory cultural mapping approach based on collective behavior in LBSNs. First, we collect the participatory sensed user behavioral data from LBSNs. Second, since only local users are eligible for cultural mapping, we propose a progressive "home" location identification method to filter out ineligible users. Third, by extracting three key cultural features from daily activity, mobility and linguistic perspectives respectively, we propose a cultural clustering method to discover cultural clusters. Finally, we visualize the cultural clusters on the world map. Based on a real-world LBSN dataset, we experimentally validate our approach by conducting both qualitative and quantitative analysis on the generated cultural maps. The results show that our approach can subtly capture cultural features, and generate representative cultural maps that well correspond with the traditional cultural maps based on survey data.

Categories and Subject Descriptors: J.4 [**Computer Applications**]: Social and Behavioral Sciences

General Terms: Design, Human Factors, Experimentation

Additional Key Words and Phrases: Cultural Mapping, Cultural Difference, Collective Behavior, Participatory Sensing, Location Based Social Networks

ACM Reference Format:
ACM Trans. Intell. Syst. Technol. 0, 0, Article 0 (0), 23 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Culture plays an important role in human evolution. It shapes both people belief system and practical behavior, which further solidify and evolve the culture. In the

This work is supported by the Swiss National Science Foundation under grant number PP00P2_153023, and the Microsoft collaborative research grant.
Authors' addresses: D. Yang, eXascale Infolab, University of Fribourg, Fribourg, Switzerland, email: dingqi.yang@unifr.ch; D. Zhang, Institut Mines-Télécom/Télécom SudParis, Evry, France, email: daqing.zhang@it-sudparis.eu; B. Qu, University of Rennes 1 – IRISA & Inria Rennes, France, email: bqu@ina.fr.

DOI: http://dx.doi.org/10.1145/0000000.0000000

long history of human development, there have been literally thousands of cultures on Earth, which differ in various aspects such as moral values, religious beliefs, language, clothing, cuisine, recreation, architecture, music and dance, etc. When studying culture, one of the primary tasks is to understand cultural difference across the world, which is valuable in many fields and can then enable various applications.

In order to identify and analyze cultural difference across the world, the United Nations Educational, Scientific and Cultural Organization (UNESCO) uses *Cultural Mapping* [Poole 2003] as a crucial tool and technique to visualize cultural differences and boundaries on the map. Basically, the goal of *Cultural Mapping* is to create the map representation of different cultures and their boundaries, from the perspectives of *indigenous and local people* with respect to various cultural aspects. In the context of cultural mapping, "cultures" are represented by cultural clusters of specific geographical areas, such as countries, regions or cities in the world. For example, Heatwole [Heatwole 2006] created a world cultural map based on the people's religious beliefs in different regions across the world. Inglehart et al. [Inglehart and Welzel 2010] built a cultural map of 53 countries based on the people's moral values extracted from the World Values Survey[1] (WVS).

Traditional cultural mapping approaches usually encompass a wide range of activities in data collection, which is mainly achieved via large-scale surveys (i.e., questionnaires and interviews) about the moral values and beliefs of the participants. However, cultural data collection via large-scale surveys usually incurs a significant cost of both human resources and time. For example, the current wave of the WVS was carried out from 2010 to 2014, involving more than 60 countries with over 1000 participants from each country.

In this study, in order to explore an efficient approach for cultural mapping, we resort to Location Based Social Networks (LBSN) for cheaply collecting the participatory sensed collective behavior data, which massively implies culture information. Specifically, since various definitions of culture suggest that collective human behavior is tightly associated with human culture [Hoebel 1972; McGrew 1998; Taylor 1967], we are motivated to tackle cultural mapping problem from collective human behavioral perspective. In order to select collective behavior data, we need to define an appropriate granularity for our culture study, i.e., the group of people considered to have the same culture. Different from the existing cultural mapping works that are mainly conducted with country granularity, we focus on analyzing the collective behavior with city granularity, because culture usually extends beyond country borders. In order to obtain the collective behavior data in a city, we resort to the participatory sensed user activity data in LBSNs. LBSNs provide users with opportunities to share their real time presence with their friends by checking in at a Point of Interest (POI), such as a French restaurant or a bar, along with a short check-in message associated with their current status. By interacting with LBSNs, users generate a significant volume of check-in data online. For example, Foursquare[2], one of the most popular LBSNs, had over 5 billion check-ins with more than 45 Million users globally by January 2014[3]. Based on the above discussion, our objective of cultural mapping is to study the cultural clusters of cities by extracting cities' cultural features from the large-scale collective behavior data in LBSNs.

However, it is not straightforward to identify useful and representative features for cultural mapping from such large-scale collective behavior data in LBSNs. By studying

[1] http://www.worldvaluessurvey.org
[2] https://foursquare.com
[3] https://foursquare.com/about

the specific characteristics of user activity data in LBSNs and the cultural implications on collective behaviors, we identify three key cultural features as follows:

— First, check-ins at POIs in a city imply the daily activity pattern in that city. Specifically, in LBSNs, a POI is usually associated with a category (e.g., French restaurant), which can be considered as the semantic representation of user activities. By analyzing such semantic information of collective check-ins in cities, we can reveal the cultural difference between cities with regard to their daily activity pattern. For example, there are obvious differences between western cuisine and oriental cuisine [Counihan and Van Esterik 2012], which leads to the difference between users' eating behavior in Paris and in Hong Kong, i.e., local users in Paris may frequently go to French or Italian restaurants while local users in Hong Kong may frequently go to Chinese restaurants or Sushi bars.
— Second, check-ins capture inter-city crowd mobility patterns, which can reflect the cultural similarity between cities. Specifically, by analyzing collective check-ins on a global scale, we can discover crowd mobility between cities. Since human are the primary carrier of culture, human mobility and migration, which are basic cultural exchange activities, are fundamental ways of cultural diffusion [Perreault and Brantingham 2011]. Therefore, crowd mobility between cities can reflect the cultural similarity between cities. Intuitively, cities with more similar cultures probably have more communication among them (i.e., a larger number of users traveling among the cities), and vice versa.
— Third, check-in messages in cities, which contain explicitly user status, imply linguistic characteristics of the cities. Concretely, as a form of human behavior, language is the principal means in human communication [Sapir 1927]. It expresses, embodies and symbolizes human culture and can thus reflect cultural differences [Kramsch 1998]. In order to understand the language usage in LBSNs in a specific city, we can conduct language detection of check-in messages. By comparing the language usage between cities, we may discover the cultural difference between cities with regard to the linguistic aspect.

In addition, although check-ins in LBSNs contain rich cultural information, not all of them are eligible for cultural mapping. Concretely, cultural mapping suggests that only *indigenous and local people* are eligible to represent local culture [Poole 2003]. Therefore, for a specific city, check-ins generated by non-local users, who are not representative for local culture, are considered as noisy data and should thus be removed when studying the city's culture. Different from a survey where we can delicately select the local people as participants, LBSNs do not allow us to select only local participants or their check-ins in the data collection process. Therefore, for a specific city, the collected check-ins often include non-local users' behavior which are not eligible in characterizing the city's culture and should thus be eliminated for cultural mapping.

In this paper, aiming at building a cultural map from human behavior perspective, we propose a participatory cultural mapping approach, based on collective behavior data in LBSNs. Specifically, the proposed approach consists of four steps. First, in order to collect large-scale user behavioral data, we collect check-ins in LBSNs on a global scale. Second, in order to detect the local users of a city, we propose a progressive "home"[4] location identification method which searches for a user's most frequented region and progressively narrows the region down to a small area. Third, by extracting the three key cultural features from local users' check-ins, i.e., daily activity pattern, inter-city mobility and linguistic feature, we propose a cultural clustering method

[4]By "home" location of a user, we mean the location around where most of the user's activities happened rather than the actual home of the user.

that builds an affinity matrix between cities based on the extracted features and then leverages spectral clustering techniques to discover the cultural clusters. Finally, we generate a cultural map by visualizing the detected cultural clusters on the world map.

We experimentally evaluate the proposed approach based on a large-scale check-in dataset collected from Foursquare. Specifically, we first conduct qualitative analysis on the cultural maps created using individual features and their combination. The results show that the proposed approach can efficiently capture cultural information from user check-in data and generate representative cultural maps. We then quantitatively compare our cultural maps with those created by the traditional cultural mapping approaches based on survey data. The results show that our cultural clusters well correspond with the traditional cultural clusters. We also observe some interesting differences caused by some unique cultural features extracted from user behavioral data.

The rest of the paper is organized as follows. We present the related work in Section 2. The overview of the proposed participatory cultural mapping approach is presented in Section 3. We explain the data collection process in Section 4. The local user identification method is presented in Section 5, followed by the cultural clustering method in Section 6. Experimental evaluation is shown in Section 7. We discuss the limitation of our work in Section 8, and conclude our work in Section 9.

2. RELATED WORK

In thousands of years of human development, there have been thousands of cultures on Earth, which lead to cultural diversity around the world. On the one hand, cultural diversity can benefit human development. For example, different cultures usually imply different ways of thinking and solutions to problems, which is an important source of creativity. On the other hand, cultural diversity may also be a barrier in human development. For example, in the context of globalization and economic development, the lack of cultural understanding has often backfired, resulting in ineffective projects and wasted investments. Therefore, it is crucial to understand cultural differences across the world. In current literature, cultural differences have been widely studied from various domains, such as psychology [Cole and Bruner 1971], genetics [Laland et al. 2010], behavior [Triandis 1989], education [Hofstede 1986], economy [Du Gay and Pryke 2002] and business management [Moran et al. 2007], etc.

In order to analyze cultural differences, UNESCO uses *Cultural Mapping* [Poole 2003] to identify and visualize cultural differences on the map. In current literature, most of the existing works focus on cultural mapping from the psychological perspective and its applications. For example, Schwartz [Schwartz 2004] investigated cultural differences from the value orientation perspective. Bond et al. [Bond and Leung 2009] studied the cultural mapping based on human beliefs and its application to a social psychology involving culture. The department of communications and the arts in Australia [of Communications and the Arts 1995] studied the cultural and economic development using cultural mapping. Evans et al. [Evans and Foord 2008] applied culture mapping on arts facilities and activity planning in the UK. However, these works not only encompass an expensive data collection process via large-scale surveys, but also fail to consider people's practical behavior which is also an important factor in culture [Skinner 1953]. Nevertheless, it is difficult to collect large-scale human behavior data in practice.

Fortunately, with the recent boom of Location Based Social Networks, users have reported a considerable number of activities (i.e., check-ins) online [Yang et al. 2015]. By analyzing user check-in data in LBSNs, we are able to explore large-scale user behavior from various perspectives. From the activity category perspective, various works consider POI categories as the semantic representation of user activities. Noulas et al.

[Noulas et al. 2011] clustered and annotated regions in a city characterized by users' activity categories. Preoţiuc-Pietro et al. [Preoţiuc-Pietro et al. 2013] explored activity category based city-to-city similarity measures, which considered a city as a bag of POI categories. Wang et al. investigated community detection in LBSNs based on user activity categories [Wang et al. 2014], and characterize cities based on community profiling [Wang et al. 2012]. From the mobility perspective, Cheng et al. [Cheng et al. 2011] studied user mobility patterns in the US with regard to geographical, economic and social factors. Yang et al. [Yang et al. 2015] studied the spatio-temporal patterns of user activities in LBSNs. Noulas et al. [Noulas et al. 2012] investigated the universal user mobility patterns in LBSNs using check-in data collected from different cities around the world. User mobility patterns can then be exploited to enable various applications, such as next location prediction [Lian et al. 2015] and offline event marketing [Yu et al. 2015]. From the linguistic perspective, Bauer et al. [Bauer et al. 2012] discovered the dominant topics in the neighborhoods of a city by applying topic modeling techniques on a collection of check-in messages. Yang et al. conducted sentiment analysis on user-left tips in LBSNs in order to augment personalized location recommendations [Yang et al. 2013a] and searches [Yang et al. 2013b].

In addition, this large-scale user behavioral data in LBSNs massively implies cultural related information, which provides us with novel opportunities of studying cultural differences. From sociological perspective, Mische [Mische 2011] discussed the possibility of mapping cultural elements based on social network data. Hemmersam et al. [Hemmersam et al. 2014] studied the opportunity of exploring locative media for cultural mapping. Golder et al. [Golder and Macy 2011] studied the cultural differences of the collective moods by applying sentiment analysis on user messages in Twitter[5]. Park et al. [Park et al. 2013] investigated the cultural differences on the usage of facial expressions for emotion in Twitter. Yuan et al. [Yuan et al. 2013] explored both individual and community lifestyles in China using user digital footprints left in LBSNs and other social networks. Silva et al. [Silva et al. 2014] studied large-scale city dynamics and identified several cultural differences on eating habits across different cities. In this paper, we propose a participatory cultural mapping approach based on the collective behavior in LBSNs. Different from existing works that mainly focus on a specific aspect of user behavior, such as linguistics [Golder and Macy 2011] or daily activity patterns [Yuan et al. 2013; Silva et al. 2014], we conduct in-depth analysis on user behavioral data from various aspects and extract three unique features (i.e., daily activity pattern, inter-city mobility and linguistic) to perform cultural clustering on a global scale.

3. OVERVIEW OF THE PARTICIPATORY CULTURAL MAPPING APPROACH

Figure 1 illustrates the overview of the proposed participatory cultural mapping approach that consists of four parts. First, we collect check-in data from LBSNs on a global scale, which capture large-scale user behavior around the world. Second, using a progressive "home" location identification method, we identify local users in a city, whose behavior is considered to be representative in characterizing the city's culture. Third, by extracting three key features from check-in data of local users in cities, we build an affinity matrix and leverage spectral clustering techniques to discover cultural clusters of cities around the world. Finally, we plot a cultural map by simply visualizing the detected cultural clusters on the world map.

[5]https://twitter.com

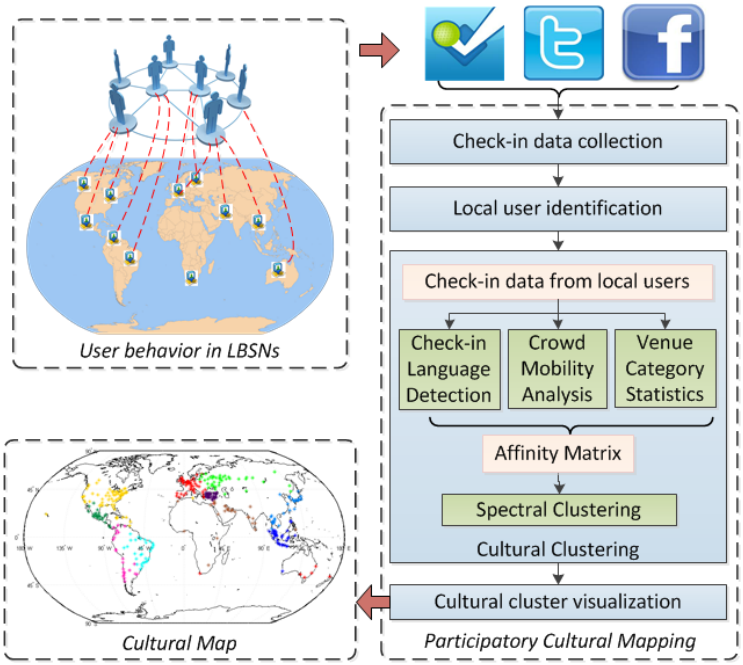


Fig. 1. Overview of the participatory cultural mapping approach

4. CHECK-IN DATA COLLECTION

In this work, we collect check-in data in Foursquare which captures large-scale user behavior around the world. Specifically, since a user's check-in information in Foursquare can only be accessed from her own social circle, they are not available publicly. However, Foursquare users can choose to post their check-ins via Twitter when they check in at a place. Hence, we capture check-ins by crawling foursquare-tagged tweets from the Twitter Public Stream[6]. Using this approach, we collected a Foursquare check-in dataset over about 18 months (from April 2012 to September 2013).

Similar to our previous work [Yang et al. 2013a], we process the raw data with various noise filtering steps. First, although Foursquare tries to verify whether a user is actually at the place when they check in there, fake check-in data is inevitable in large datasets. We observed that some users had performed "sudden-move" check-ins (consecutive check-ins with a speed faster than 1200 km/h: the normal speed of an airplane). Therefore, all the check-ins from these "sudden-move" users are eliminated. These "sudden-move" users represent about 1.1% of all the users in our dataset, while their check-ins represent about 3.4% of all the check-ins in the collected data. Second, some of the venues[7] in our dataset cannot be resolved by Foursquare venue API, causing the venue category information of these venues to be unavailable. Since venue category is critical to semantically understand user behavior, we also excluded the check-ins performed at these venues, which are about 7.5% of all the venues. The check-ins at these venues only represent less than 1.0% of all the check-ins in the collected data because these venues are usually unpopular. Third, we select only check-ins from active

[6]https://dev.twitter.com/docs/streaming-apis/streams/public
[7]Since "venue" is used to represent a POI in Foursquare, we don't differentiate these terms throughout this paper.

users (defined as users who have performed at least one check-in per week). After noise filtering, our dataset includes 279,495 users who have performed 49,273,956 check-ins at 6,743,711 venues globally.

5. IDENTIFICATION OF LOCAL USERS

Cultural mapping suggests that only local users' activities in a city are eligible to characterize the culture of the city. In order to identify local users in a city, we need to know the home location of each user. However, due to privacy protection, such information cannot be accessed from Foursquare. Moreover, although Twitter gives users the option to register a home location for their accounts, only a limited number of users provide valid information. Therefore, it is necessary to algorithmically identify the home location for each user.

Intuitively, we can simply search for a small area where a user checks in most frequently and regard the center of this area as her home location. However, directly searching such a small area from a mass of check-ins may overlook some user-frequented but relatively-large areas, which then leads to inappropriate home location identification. For example, considering a New York user who frequently goes to Boston for business trips and has a high check-in frequency at a few POIs there (e.g., the office of a business partner and a nearby hotel), the identified home location may probably be in Boston although the check-ins are massively around New York and its surrounding area.

In this paper, rather than directly searching for a small area to identify a user's home location, we propose a progressive home location identification method. For a specific user, it starts from searching for a large region where most of the user's check-ins happen, and then repeat the search with a reduced region size within the large region, until a small region is identified. In current literature, a similar approach [Cheng et al. 2011] has been used to identify user's home location in Twitter, which first segments an area into disjoint grid cells and then recursively searches for the most-checked grid cell with the decreasing cell size. However, this method may cause inaccuracy due to the segmentation process, particularly when a densely checked area is segmented into several disjoint grid cells. Therefore, instead of searching for disjoint grid cells, we iteratively search the circular regions (with a certain radius) centered by each user checked venue and select the most checked region. By repeating this step with the decreasing radius, we finally obtain a small region where the user checks in most frequently and we regard the center of this region as the user's home location. Figure 2 illustrates an example of the progressive home location identification method.

Formally, for a specific user u, we denote her check-ins as A_u which contains the set of the checked venues V_u. Each venue v is associated with a physical location $v.l$ (represented by GPS coordinates). In order to identify a user's home location, the proposed method aims to find a circular region $r_{l,d}$ with center l and radius d where the user has checked in most frequently. It contains a key step, i.e., the most-frequented region search, which is to search the most-frequented region $r_{l,d}$ in a given search region R. We then repeat this step with a decreasing region radius until a small region is found.

Algorithm 1 presents the most-frequented region search process. The basic idea is to iterate all the user checked venues in the search region R and count its surrounding check-ins in order to find the most checked region. Specifically, given a search region R and a target region radius d, we first get the venues in R where the user has ever checked in, denoted as $V_{u,R}$ (Line 1-2). Afterward, for each $v \in V_{u,R}$, we calculate the number of the user's check-ins in a candidate region r with center $v.l$ and radius d, denoted as $|A_{u,r}|$ (Line 3-6). Finally, we select the region r where $|A_{u,r}|$ is maximum (Line 7).

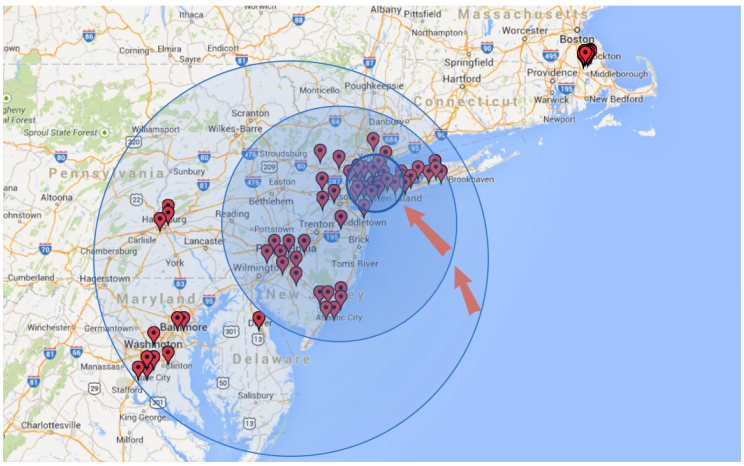


Fig. 2. Example of the progressive home location identification method (Map image courtesy of Google)

ALGORITHM 1: Most-Frequented Region Search

Data: User u's check-ins A_u and search region R, target region radius d
Result: Most frequented region r_{freq}
1 Get the set of user checked venues V_u
2 Select V_u in the search region R, denoted as $V_{u,R}$
3 **for** $v \in V_{u,R}$ **do**
4 | Select a region r with center $v.l$ and radius d
5 | Count u's check-ins in r, denoted as $|A_{u,r}|$
6 **end**
7 Get $r_{freq} = argmax_r |A_{u,r}|$

In order to identify a user's home location, we repeat the most-frequented region search process and recursively look for a smaller region in the larger region identified from the previous iteration. Algorithm 2 presents the progressive home location identification method. The algorithm requires a predefined target region radius d for each iteration. We denote the set of the predefined target region radiuses as D. Given a user's check-ins A_u and a set of target region radius D, since we start from looking for the large most-frequented region on a global scale, we initialize the search region R as the global scale (Line 1), and sort target radiuses in D in descending order (Line 2). For each target region radius $d \in D$, we search the most-frequented region r with radius d in R (Line 4), and then set the next search region to the identified r (Line 5). At the end of the iteration, we return the center of the smallest most-frequented region as the user's home location (Line 7).

In this work, since we focus on identifying a user's home location at city granularity, we empirically select a set of radius $D = \{50km, 5km, 0.5km\}$, and perform home location identification. By randomly checking 500 users who have reported their home location information in Twitter, we find that there are 75% of the users (i.e., 375 users) who report valid home locations (such as GPS coordinates, specific address, or city names, etc.) that can be resolved by Google Maps[8] to get the related city information. By verifying the identified home location with the user reported city information, our method achieves an accuracy of 88.53% (i.e., 332 users' home cities are correctly identi-

[8]https://maps.google.com

ALGORITHM 2: Progressive Home Location Identification

Data: User u's check-ins A_u, a set of target region radius D
Result: Home location l

```
Initialize the search region R as the global scale
Sort D in descending order
for d ∈ D do
    Search the most-frequented region r with radius d in R by Algorithm 1
    Set the next search region R = r
end
Get the center l of r as the user's home location
```

fied). Compared to directly searching for the small region (i.e., $d = 0.5km$) which results in an accuracy of 72.27% and recursively searching disjoint grid cells [Cheng et al. 2011] which results in an accuracy of 83.47%, the proposed progressive home location identification method achieves the best performance. More sophisticated methods (e.g., considering the text content) may be used to improve the performance. However, since it is not the main focus of this work, we use the proposed progressive home location identification method to identify the local users of a specific city.

6. CULTURAL CLUSTERING

By analyzing the collective behavior of the local users in cities, we can study the cultural differences between cities and discover cultural clusters based on these differences. Specifically, we first extract three key cultural features from check-in data, i.e., daily activity pattern, inter-city mobility and linguistic feature, in order to quantitatively measure cultural similarity and build an affinity matrix of cities. We then leverage spectral clustering techniques to discover cultural clusters based on the built affinity matrix.

6.1. Feature Extraction

User check-ins in LBSNs massively imply cultural information. First, the daily activity pattern in a city can be characterized by the categories of user checked POIs. Second, the inter-city mobility representing cultural exchange activities can be extracted from check-ins. Third, the linguistic feature characterized by the practical language usage can be obtained from check-in messages by leveraging language detection techniques. For each feature, we quantitatively measure the cultural similarity between cities.

6.1.1. ***Daily Activity Pattern.*** By collecting the check-ins of local users in a city, we are able to understand the citizen's collective daily activities. POI categories can be regarded as the semantic representation of users' activities when checking in. For example, checking in at a French restaurant probably means the user is having French food there. Therefore, we characterize a city's daily activity pattern by the collective check-in distribution on different POI categories.

Venues in Foursquare are organized with a three-level hierarchical category classification[9] by the date of data collection. Specifically, it contains 9 root categories (i.e. Arts & Entertainment, College & University, Food, Great Outdoors, Nightlife Spot, Professional & Other Places, Residence, Shop & Service, Travel & Transport) which are further classified into 291 categories at the second level. Moreover, a few second-level categories have sub-categories at the third level. Due to the incompleteness of third-level categories, only a few venues have the category information at third level.

[9]https://developer.foursquare.com/docs/venues/categories

(a) New York (b) Tokyo

Fig. 3. Examples of venue category tag clouds (Larger font size implies higher frequency, and vice versa.)

Therefore, we choose to use the second-level categories (291 categories) to semantically characterize users' behavior when checking in at POIs. Figure 3 demonstrates venue category tag clouds in New York and Tokyo. We first observe that offices, subway and train stations are popular check-in POIs in both cities. Most importantly, we observe clearly the cultural differences between the two cities, i.e., New York users usually check in at home, bars, gyms, outdoor places while Tokyo users often check in ramen/noodle houses, convenience stores, and Japanese restaurants.

Using the second level venue categories provided by Foursquare, we characterize a city's daily activity pattern using a 1×291 vector, representing the check-in distribution on the 291 venue categories. In order to quantitatively measure the difference on daily activity pattern between cities, we resort to the Jensen-Shannon divergence[Lin 1991] to measure the difference between two probability distributions because it is a *symmetric* and *bounded* metric. Specifically, given two distributions P_1 and P_2, the Jensen-Shannon divergence is calculated as follows:

$$JSD(P_1||P_2) = \frac{1}{2}KLD(P_1||M) + \frac{1}{2}KLD(P_2||M) \tag{1}$$

where $M = \frac{1}{2}(P_1 + P_2)$ and $KLD(P||M)$ is the Kullback-Leibler divergence [Kullback and Leibler 1951] which is calculated as:

$$KLD(P||M) = \sum_i \log_2(\frac{P(i)}{M(i)}) \cdot P(i) \tag{2}$$

We see that the Jensen-Shannon divergence can be regarded as the symmetrized and smoothed version of the Kullback-Leibler divergence. Using the base 2 logarithm, the Jensen-Shannon divergence is bounded in $[0, 1]$ [Lin 1991]. Therefore, for two specific cities C_1 and C_2, given their check-in distribution on venue categories P_{C_1} and P_{C_2}, we define the daily activity pattern similarity between the two cities, denoted by Sim_{DAP}, as follows:

$$Sim_{DAP}(C_1, C_2) = 1 - JSD(P_{C_1}||P_{C_2}) \tag{3}$$

6.1.2. ***Inter-city Mobility.*** By quantitatively analyzing check-ins in different cities, we are able to understand the inter-city mobility, which implies the inter-city cultural similarity. Intuitively, users in two cities that share similar culture will probably be easy to communicate and interact (e.g., doing business) among them, and thus probably have more travels from one to the other. Therefore, we investigate the behavior of users who have ever checked in in multiple cities. In order to calculate the similarity between two cities in terms of inter-city mobility, we adopt Jaccard coefficient to measure the ratio of users who have ever checked in the two cities.

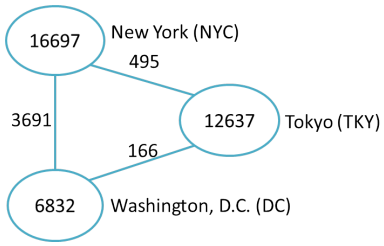


Fig. 4. An example of inter-city mobility among New York, Washington D.C. and Tokyo

Specifically, for two given cities C_1 and C_2, we denote the users who have checked in each of them as U_{C_1} and U_{C_2}, respectively. We then calculate the similarity between C_1 and C_2 in terms of inter-city mobility, denoted by Sim_{Mob}, using Jaccard coefficient.

$$Sim_{Mob}(C_1, C_2) = \frac{|U_{C_1} \bigcap U_{C_2}|}{|U_{C_1} \bigcup U_{C_2}|} \tag{4}$$

Figure 4 presents an example of three cities in our dataset, i.e., New York, Washington D.C. and Tokyo. The number inside the circle presents the city's total user number. The number on the link between cities presents the number of users who have checked both of the cities. We then calculate the similarity between them based on the inter-city mobility as follows.

$$Sim_{Mob}(NYC, DC) = 0.1861 \tag{5}$$

$$Sim_{Mob}(NYC, TKY) = 0.0172 \tag{6}$$

$$Sim_{Mob}(DC, TKY) = 0.0086 \tag{7}$$

Due to the cultural differences between Japan and U.S., we observe that the similarity with respect to user mobility between New York and Washington D.C. is significantly higher than that between Tokyo and New York (or Washington D.C.).

6.1.3. ***Linguistic Feature.*** Check-in messages massively imply the linguistic characteristics of a city, which play an important role in human culture. In this work, by conducting analysis on check-in messages, we investigate the practical language usage in LBSNs in cities. Specifically, by applying language detection techniques on check-in messages in a city, we can characterize the linguistic feature of a city by the distribution of the languages used in the city, and then quantitatively measure the similarity between cities. Due to the complexity and difficulty of multilingual text analysis (e.g., multilingual sentiment analysis), which is also beyond our focus, we do not explore the content and the detailed semantic meaning of check-in messages in this work.

In order to identify the language of a check-in message, we leverage the language detection library developed by Cybozu Labs [Shuyo 2010], which has been successfully used in processing user comments in social media [Yang et al. 2014]. We detect 47 languages in total (including "other" for unknown languages) in our dataset. Figure 5 demonstrates the top 10 languages and their percentages in our dataset. We find that English is the dominant language used in LBSNs. Some popular languages, such as Spanish and Portuguese, also appear at the top of the list. Similar results have also been reported in [Leetaru et al. 2013]. By excluding the unknown languages, i.e., "other", we can characterize a city by a distribution of check-ins on 46 languages. Figure 6 illustrates two tag clouds of languages in two big cities, i.e., Mexico City and Rio de Janeiro. We observe that English is the most popular language in both cities in LBSNs, even though it is not the official language in either of the cities. This is due

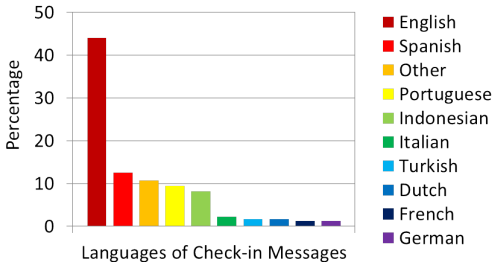


Fig. 5. Top 10 languages of check-in messages

(a) Mexico City (b) Rio de Janeiro

Fig. 6. Examples of language tag clouds

to the fact that the languages in LBSNs are biased towards English. However, we can still discover the linguistic difference between the cities, i.e., Spanish and Portuguese are the second most popular languages in Mexico City and Rio de Janeiro, respectively.

Similar to the daily activity pattern, we leverage the Jensen-Shannon divergence to measure the linguistic difference between cities. Formally, for two specific cities C_1 and C_2, we denote their distributions of check-ins on languages as L_{C_1} and L_{C_2}, respectively. We define the linguistic similarity between the two cities, denoted by Sim_{Lin}, as follows:

$$Sim_{Lin}(C_1, C_2) = 1 - JSD(L_{C_1}||L_{C_2}) \tag{8}$$

6.1.4. **Affinity Matrix Construction**. By characterizing the culture similarity from three different aspects, i.e., daily activity pattern, inter-city mobility and linguistic feature, we combine them into a unique measure by leveraging their geometric mean.

$$Sim = \sqrt[3]{Sim_{DAP} \cdot Sim_{Mob} \cdot Sim_{Lin}} \tag{9}$$

It ensures that two cities are similar if and only if they are similar in all three aspects. Since all the similarity measures, i.e., Sim_{DAP}, Sim_{Mob} and Sim_{Lin}, are bounded in $[0, 1]$, it is easy to prove that Sim is also bounded in $[0, 1]$. For a given set of cities, by calculating all the similarities between each pair of cities, we can then construct an affinity matrix in order to discover cultural clusters from it.

6.2. Spectral Clustering

Given an affinity matrix measuring the cultural similarity between cities, we adopt the spectral clustering techniques [Von Luxburg 2007], which are widely adopted in various clustering problems due to the quality of the clusters generated and the simplicity of implementation. We use a variation of spectral clustering proposed in [Ng et al. 2002] which integrates a normalization step and shows better performance compared to the classical spectral clustering algorithm [Von Luxburg 2007]. Moreover, similar

ALGORITHM 3: Spectral Clustering with Auto-Selected Number of Clusters

Data: Affinity matrix M, Range of the number of clusters $[k_{min}, k_{max}]$
Result: Clusters $\{C_1, ..., C_k\}$
1 Construct the diagonal degree matrix D that $D(i,i) = \sum_{j=1}^{n_c} M(i,j)$
2 Calculate the Laplacian matrix $L = D - M$
3 Calculate the normalized Laplacian matrix $L_{norm} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$
4 Get the k_{max} smallest eigenvalues $\{\lambda_1, ..., \lambda_{k_{max}}\}$ of L_{norm}
5 Calculate $\delta_i = \lambda_{i+1} - \lambda_i$ for $i \in [k_{min}, k_{max} - 1]$
6 Select the optimal $k = argmax_i \delta_i$
7 Get the k smallest eigenvectors $\{e_1, ..., e_k\}$ of L_{norm}
8 Construct a matrix X where e_i is its ith column
9 Treat each row of X as a data sample and cluster them into k clusters using k-means, denoted as $\{C_1, ..., C_k\}$

to [Cranshaw et al. 2012], we also integrate a method to auto-select the number of clusters within a given range.

Algorithm 3 presents the clustering process. Let M denote an affinity matrix of cities, with the size of $n_c * n_c$, where n_c is the number of cities. We also define a range for the number of clusters, denoted as $[k_{min}, k_{max}]$, as the inputs. In order to conduct spectral clustering, we start by calculating the normalized Laplacian matrix (Line 1-3). We then select the optimal number of clusters k in $[k_{min}, k_{max}]$ by searching for the largest gap between two consecutive eigenvalues (Line 4-6). Finally, by calculating the k smallest eigenvectors and use them to represent the data samples, we adopt k-means to cluster them into k clusters (Line 7-9). Please refer to [Ng et al. 2002] for more mathematical details about spectral clustering.

Once we obtain the cultural clusters, combined with the location of the cities, we create a cultural map by visualizing these clusters using different colors on the map.

7. EXPERIMENTAL EVALUATION

In order to validate the proposed participatory cultural mapping approach, we carry out various experiments based on the large-scale check-in data collected from Foursquare, and conduct both in-depth qualitative and quantitative analysis on the generated cultural maps. Specifically, by selecting 415 big cities around the world, we qualitatively study the cultural map generated using the proposed approach and the implication of the individual features, and show some interesting cultural correlations between user behavior and other factors such as geography, immigration, religion, etc. Moreover, by comparing the cultural maps (or cultural clusters) created using the survey data from the WVS [Inglehart and Welzel 2010] and the GLOBE[10] (Global Leadership and Organizational Behavior Effectiveness research) project [Gupta et al. 2002], we quantitatively evaluate the proposed approach and discuss its advantages and limitations.

7.1. Dataset Selection

In this work, since we focus on the cultural map with city granularity, we leverage a dataset of world cities provided by ESRI[11], a leading company in Geographic Information System (GIS). The dataset consists of 2535 major cities around the world, most of which are national or provincial capitals. Combined with the check-in dataset, we plot the cumulative distribution of the number of check-ins in cities in Figure 7. We observe

[10]http://www.tlu.ee/širvir/Leadership/Leadership%20Dimensions/globe_project.html
[11]http://www.esri.com/

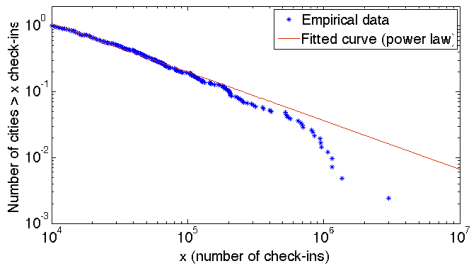


Fig. 7. Cumulative distribution of the number of check-ins in cities

Table I. Statistic of the Selected Dataset

Number of valid cities	415
Number of countries	77
Number of check-ins	33,263,233
Number of users	266,909
Number of venues	3,680,126

Fig. 8. Tag cloud of countries where the selected cities are located

that it follows a power-law distribution [Clauset et al. 2009], i.e., $P(x) \sim x^{-\beta}$ with the estimated $\beta = 0.74$, which implies that there are a large number of cities with a small number of check-ins. Intuitively, the user check-ins in these less checked cities may not be sufficient or representative enough to characterize the cities' culture. Therefore, in order to filter out the less checked cities, we select the cities containing more than 10,000 check-ins as valid cities, resulting in 415 valid cities located in 77 countries. Table I presents the statistics of the selected dataset. Figure 8 illustrates the tag cloud of countries where these 415 cities are located. Unsurprisingly, the United States, where Twitter and Foursquare started their business, has most cities (i.e., 60 cities) in the dataset.

7.2. Qualitative Evaluation

In order to qualitatively evaluate the proposed approach, in this section, we first analyze the cultural map created with the selected dataset, and then discuss the implications of the individual features (i.e., daily activity pattern, inter-city mobility and linguistic feature) in cultural mapping. Based on the cultural maps/clusters presented in [Inglehart and Welzel 2010; Heatwole 2006] and [Gupta et al. 2002], where the number of clusters are 9, 10 and 13, respectively, we empirically set the range for the number of cultural clusters as $[5, 15]$ and let Algorithm 2 select the optimal one in this range.

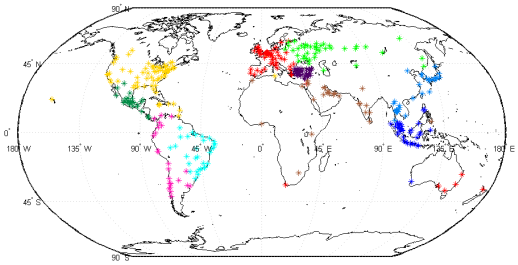


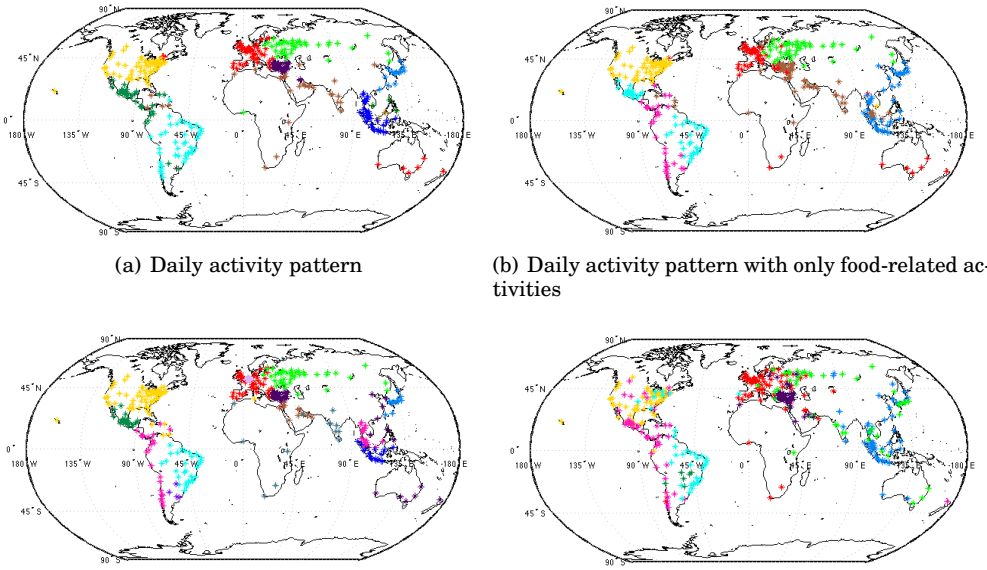
Fig. 9. Cultural map created by the proposed approach

Figure 9 demonstrates the cultural map created by our approach, which identifies 10 clusters in total. First, we observe the geographical constraint on cultural clusters, i.e., cities in a cluster are located in a specific region in the world. For example, some major cultural clusters are North America, Middle America, South America, Western Europe, Eastern Europe, the Middle-East, East Asia and South Asia. However, the geographical proximity between two cities does not necessarily mean that they are associated with the same cluster. For example, Cape Town in South Africa and some cities in Australia and New Zealand are associated with the Western Europe cluster, even though they are geographically distant from Western Europe. It is probably due to the emigration from the UK to these cities which were its colonies in the past. Second, we observe that there are two clusters in Latin America, which are separately located in the Eastern part and Western part of Latin America. It is probably due to the language usage in Latin America, i.e., Portuguese is the dominant language in West Latin America while Spanish is the most popular language in East Latin America. Third, cities in Turkey, Greece and Cyprus form a standalone cluster surrounded by Western Europe, Eastern Europe and Middle-East clusters. By investigating the individual features of these cities, we find that there is a significant mobility among these cities, which then leads to a standalone cluster in this area.

In order to further evaluate our approach, we create cultural maps based on individual features and then study the implications of these features in cultural mapping. Figure 10 presents four cultural maps based on the cities' daily activity pattern, daily activity pattern with only food related activities, inter-city mobility and linguistic feature. Note that the colors of the clusters are assigned in the way that they can be better visually distinguished, and there is not necessarily a correspondence between the clusters with the same color in different cultural maps.

7.2.1. ***Daily activity pattern.*** Figure 10(a) demonstrates the cultural map created based on the daily activity pattern in cities, where 9 cultural clusters are identified. Although it looks similar to the cultural map created using all the features, there are still some interesting differences. First, only one cluster dominates Latin America due to the similar activity patterns in the cities there. Second, Montreal, which was a French colony for over 200 years and is now a bilingual (i.e., French & English speaking) city in Canada, is clustered together with Western European cities. By investigating the check-in POI categories in Montreal, we find that there are a significant number of French Restaurants there, which is the primary reason that it is put in the Western Europe cluster.

Furthermore, since food is a fundamental element in a culture [Counihan and Van Esterik 2012], the cultural difference of food has been studied from various perspectives, such as flavor [Ahn et al. 2011] and recipe [Zhu et al. 2013]. Therefore, we are motivated to study the food preferences across the world by analyzing people's eating



(a) Daily activity pattern

(b) Daily activity pattern with only food-related activities

(c) Inter-city mobility

(d) Linguistic feature

Fig. 10. Cultural maps based on individual features (Note that the colors of the clusters are assigned in the way that they can be better visually distinguished, and there is no strict correspondence between the clusters with the same color in different cultural maps.)

activities reported to LBSNs. Specifically, we create a cultural map only based on the food-related activities (i.e., check-ins at the POIs of the "Food" root-category including 88 sub-categories). Figure 10(b) presents the cultural map based on the food-related activities, which include 7 clusters. We observe that Montreal is still with the Western Europe cluster. Interestingly, there are a number of cities in South East Asia, particularly in Malaysia and Indonesia, which are in the same cluster as Middle-Eastern cities, although they are geographically distant. It can probably be explained by the religious similarity and its impact on the food preferences in those cities. While Islam is the largest religion in the Middle East, it is also the most widely practiced religion in Malaysia and Indonesia [Bell 2012]. Although Middle Eastern cities are geographically distant from Malaysian and Indonesian cities, due to the same Islamic dietary law, the food preferences in these cities are similar.

7.2.2. ***Inter-city mobility.*** Figure 10(c) presents the cultural map based on the inter-city mobility, where 14 clusters are identified. We observe strong geographical constraints on the clusters. First, due to the power-law distribution of travelling distance in LBSNs [Cheng et al. 2011; Noulas et al. 2012], the inter-city mobility tends to be significant within small areas, which leads to more clusters with a small geographical span. In addition, the administrative constraints (e.g., visa applications) also mean that a number of users may only travel within their own countries, which is also the reason that cultural analysis is often conducted with country granularity in current literature [Inglehart and Welzel 2010; Gupta et al. 2002].

7.2.3. ***Linguistic feature.*** Figure 10(d) presents the cultural map based on the linguistic feature, where 7 clusters are identified. We observe that the clusters do not have clear geographical boundaries between each other and thus overlap. This is due to the fact that check-in languages in LBSNs are biased towards English. Although the lan-

Table II. Cultural clusters and the related city numbers

Cultural Clusters of WVS [Inglehart and Welzel 2010]		Cultural Clusters of GLOBE [Gupta et al. 2002]	
Cluster Name	Number of cities	Cluster Name	Number of cities
English Speaking	83	Anglo	85
Catholic Europe	17	Latin Europe	15
Protestant Europe	21	Nordic Europe	3
Orthodox	39	Germanic Europe	19
Latin America	75	Eastern Europe	38
Islamic	40	Latin America	67
South Asia	57	Arab	42
Confucian	22	Southern Asia	56
		Confucian Asia	23

guages of check-in messages are biased representations of the languages in a city, we can still observe some interesting clusters. For example, Latin American is separated into two clusters due to the fact that Portuguese is the official language in Brazil, while Spanish is the most popular language in most of the other Latin American countries.

7.3. Quantitative Evaluation

Different from the traditional cultural mapping approaches that mainly collect data via large-scale surveys, we propose a participatory cultural mapping approach that leverages the participatory sensed collective behavior data. In this section, since a cultural map intrinsically consists of a set of cultural clusters, we quantitatively evaluate the proposed approach by comparing the traditional cultural clusters based on survey data and the ones generated by the proposed approach. Specifically, we first find two related works that identify cultural clusters using survey data, and then select the valid cultural clusters including the common cities in our dataset and their dataset. By applying our cultural mapping approach on the check-in data in the related cities with different features, we conduct an overall comparison between the obtained cultural clusters and the traditional cultural clusters using Normalized Mutual Information (NMI). Finally, by conducting the cluster-wise comparison, we analyze the correlation and the differences between our cultural clusters and the traditional clusters.

7.3.1. ***Traditional cultural clusters based on survey data.*** We have found two works related to cultural mapping using the survey data from the WVS and the GLOBE project in current literature. Specifically, based on the people's moral value data in 53 countries from WVS, Inglehart et al. [Inglehart and Welzel 2010] created a cultural map consisting of 9 cultural clusters, viz., "English Speaking", "Catholic Europe", "Protestant Europe", "Orthodox", "Latin America", "Africa", "Islamic", "South Asia" and "Confucian". Based on people's leadership psychology [Bass 1960] data in 61 countries from the GLOBE Project, Gupta et al. [Gupta et al. 2002] identified 10 cultural clusters, viz., "Anglo", "Latin Europe", "Nordic Europe","Germanic Europe", "Eastern Europe", "Latin America", "Arab", "Sub-Saharan Africa", "Southern Asia" and "Confucian Asia".

However, while our approach focuses on city granularity, these works all focus on the cultural clusters on country granularity. In order to bridge this gap, we consider all cities in a country to be within the same cluster and then obtain the corresponding cultural clusters on city granularity. Moreover, since our dataset contains 77 countries in total, we only use the cities that appear in both our dataset and the dataset in [Inglehart and Welzel 2010] or [Gupta et al. 2002] for comparison. As a result, due to the low popularity of LBSNs in Africa, only two cities in our dataset appear in the "Africa" cluster in the WVS dataset, and none of the cities appears in the "Sub-Saharan Africa" cluster in the GLOBE dataset. Therefore, we remove these clusters and filter out the cities concerned. Finally, we obtain 8 cultural clusters with 354 cities

for [Inglehart and Welzel 2010] and 9 cultural clusters with 348 cities for [Gupta et al. 2002]. Table II presents the cultural clusters and the number of cities in individual clusters.

We observe that the cultural clusters in these two works have an obvious correspondence. The main difference is that Gupta et al. considered countries in "Nordic Europe" as a standalone cluster, while Inglehart et al. put them in the "Protestant Europe" cluster.

7.3.2. ***Overall comparison with traditional cultural clusters.*** Based on the selected cities of WVS and GLOBE Project, we identify cultural clusters using the proposed approach with individual features as well as their combination. We then calculate the Normalized Mutual Information (NMI) [Ana and Jain 2003] between our cultural clusters and the traditional cultural clusters, which measures the correlation between them. Formally, for a dataset of N cities, we denote two sets of cultural clusters as $\Omega = \{\omega_1, \omega_2, ..., \omega_K\}$ and $\Phi = \{\phi_1, \phi_2, ..., \phi_J\}$, where ω_k represents the kth cluster in Ω, and so on. The normalized mutual information is calculated as follows:

$$NMI(\Omega, \Phi) = \frac{2I(\Omega, \Phi)}{H(\Omega) + H(\Phi)} \tag{10}$$

where I is the mutual information between Ω and Φ. $H(\Omega)$ and $H(\Phi)$ is the entropy of Ω and Φ, respectively. Using maximum likelihood estimation, they are calculated as follows:

$$I(\Omega, \Phi) = \sum_{k=1}^{K} \sum_{j=1}^{J} \frac{|\omega_k \cap \phi_j|}{N} \log \frac{\frac{|\omega_k \cap \phi_j|}{N}}{\frac{|\omega_k|}{N} \frac{|\phi_j|}{N}} \tag{11}$$

$$H(\Omega) = -\sum_{k=1}^{K} \frac{|\omega_k|}{N} \log \frac{|\omega_k|}{N}, H(\Phi) = -\sum_{j=1}^{J} \frac{|\phi_j|}{N} \log \frac{|\phi_j|}{N} \tag{12}$$

Note that N is the number of cities in the dataset. The value of NMI is actually bounded in $[0, 1]$. The higher value implies higher correlation between two sets of clusters. Please refer to [Ana and Jain 2003] for more mathematical details.

Table III presents the experiment results. First, we observe that the numbers of the cultural clusters identified by our approach are quite similar to that of traditional cultural mapping approaches. Second, the incorporation of all features in our approach results in the best NMI and outperforms all the results using individual features. Moreover, comparing the results using individual features, we find that daily activity pattern feature results in the best NMI, followed by inter-city mobility feature. Due to the bias of language usage in LBSNs, the linguistic feature yields the worst results.

We also calculate the NMI between WVS and GLOBE cultural clusters, resulting in 0.9175. Such a high value of NMI can be explained by two reasons. First, since the original cultural granularity of WVS and GLOBE is at country level, we consider all cities in a country to be within the same cluster. Such a mapping scheme brings an intrinsic consistency to WVS and GLOBE cultural clusters, while our cultural mapping approach tackles the problem directly with a finer granularity at city level. Second, WVS and GLOBE all study cultural maps from human belief and moral value perspective instead of behavioral perspective, which implies that they would have relatively high consistency.

7.3.3. ***Cluster-wise comparison with traditional cultural clusters.*** In order to better understand the difference between our cultural clusters and traditional ones, we further analyze the correlation between each pair of clusters. Specifically, for each of our clusters, we calculate its purity with respect to each traditional cultural cluster, i.e., the

Table III. Normalized Mutual Information between different sets of cultural clusters

Features	WVS data [Inglehart and Welzel 2010]		GLOBE data [Gupta et al. 2002]	
	Identified cluster number	NMI	Identified cluster number	NMI
Daily activity pattern	7	0.7446	7	0.7273
Inter-city mobility	8	0.7154	8	0.6789
Linguistic feature	9	0.5138	8	0.5354
All features	8	0.7671	7	0.7415

Table IV. Comparison of individual cultural clusters (WVS dataset)

	C1	C2	C3	C4	C5	C6	C7	C8
English Speaking	0.02	0	**0.96**	0	0	0.21	0.06	0.03
Catholic Europe	0	0	0	0	0	0.28	0	0
Protestant Europe	0	0	0	0	0.03	**0.34**	0	0
Orthodox	0	0	0	0	**0.97**	0.03	0	0
Latin America	0	**1**	0.01	0	0	0	0	**0.97**
Islamic	0	0	0	**0.92**	0	0.08	0	0
South Asia	0.41	0	0.03	0.08	0	0.06	**0.94**	0
Confucian	**0.57**	0	0	0	0	0	0	0

Table V. Comparison of individual cultural clusters (GLOBE dataset)

	C1	C2	C3	C4	C5	C6	C7
Anglo	0	0.03	**0.96**	0	0	0.22	0.1
Latin Europe	0	0	0	0	0	0.21	0
Nordic Europe	0	0	0	0	0	0.04	0
Germanic Europe	0	0	0	0	0	**0.28**	0
Eastern Europe	0	0	0	0.05	**0.86**	0.05	0
Latin America	0	**0.97**	0.01	0	0.14	0	0
Arab	0	0	0	**0.95**	0	0.1	0
Southern Asia	0.39	0	0.03	0	0	0.1	**0.87**
Confucian Asia	**0.61**	0	0	0	0	0	0.03

percentage of its cities that appear in each of tradition cultural clusters. Table IV and V present the results on the WVS and Globe dataset, respectively. Taking the first column in Table IV as an example, 2%, 41% and 57% of the cities in our cultural cluster C1 belong to the "English Speaking", "South Asia" and "Confucian" cultural clusters, respectively. We highlight the highest percentage in each column, which indicates the most relevant traditional cultural cluster for each of our clusters. Please note that due to the dataset differences between WVS and Globe, there is no strict correspondence between the clusters identifier (e.g., C1 in Table IV is not the same as C1 in Table V).

On the one hand, we observe that the cultural clusters identified by the proposed approach are highly correlated with the traditional cultural clusters. Specifically, some traditional cultural clusters can be obviously identified in both WVS and GLOBE datasets, i.e., the cities of those clusters mostly appear in only one of our clusters. For example, with the WVS dataset, "English Speaking", "Orthodox", "Islamic", "South Asia" and "Confucian" clusters are clearly associated with C3, C5, C4, C7 and C1, respectively. With the Globe dataset, "Anglo", "Eastern Europe", "Arab", "Southern Asia" and "Confucian Asia" clusters are clearly associated with C3, C5, C4, C7 and C1, respectively.

On the other hand, we also observe some interesting differences between our clusters and the traditional cultural clusters. First, with the WVS dataset, two of our clusters, i.e., C2 and C8, are associated with "Latin America" clusters. This is mainly due to the consideration of linguistic feature in our approach, i.e., Spanish and Portuguese are two major languages in Latin America. Second, with the WVS dataset, the cities in Western Europe are put together in one cluster, i.e., 21%, 28% and 34% of the cities in

cluster C6 are associated with "English Speaking", "Catholic Europe" and "Protestant Europe" clusters, respectively. A similar observation can also be found in the cluster C6 using the GLOBE dataset. By investigating the similarity between the related cities based on the individual features, we found that these cities are very similar to each other with respect to the inter-city mobility and the daily activity pattern. Therefore, our approach cannot distinguish them, and thus puts them in the same cluster. It is probably due to the fact that the LBSN user behavior data is a biased sample of collective behavior, and it is not able to reflect significant cultural differences among those cities of European Union. Such data bias will be discussed later. In contrast, these cities may probably differ in the moral value dimension, which leads to three different clusters in WVS and GLOBE.

8. DISCUSSION

Demographic bias in LBSNs. User adoption of LBSNs usually varies across different countries. For example, in our dataset, we observe that cities in less developed countries, such as many cities in Africa, have a very limited number of LBSN users, which leads to a small number of user behavioral data. In our current study, we have filtered out the cities with less check-in data, which is a common approach in social network analysis. In the future, in order to further verify the applicability of our framework, we plan to investigate more in the cities with low penetration rate of social media, and study the appropriate penetration rate to make our framework beneficial.

Behavior data bias in LBSNs. User check-ins in LBSNs may not be a full representation of users' daily activities. Since users voluntarily report their activities in LBSNs, check-ins are biased samples of user daily activities, which can be regarded as a social representation of user activities (biased towards certain social activities/POI categories, e.g., shopping and food related activities). Moreover, the user community of LBSNs may be biased towards tech-savvy and young people who prefer to use social network services. However, despite the existence of these data bias in LBSNs, our study shows that check-in data still contains valuable cultural information and can be used to generate representative cultural maps. In the future, we plan to investigate more into the influence of such data bias on the cultural mapping.

Temporal dynamics of culture and collective behavior. It is known that culture can spread from one area to another due to the various cultural exchange activities, such as immigration. In the human history, such culture diffusion process is usually slow over time and gradually leads to globalization [Robertson 1992]. In this study, due to the limited duration of user behavioral data collection process, we do not investigate the temporal dynamics of culture and collective behavior. However, as we continuously collect social media data, we believe that in the future, we can track cultural diffusion by studying long-term user activity data in LBSNs.

9. CONCLUSION

Cultural mapping has been recognized as a crucial tool by UNESCO to visualize cultural difference and culture boundaries on the map. Traditional cultural mapping approaches usually rely on large-scale survey data with respect to human belief, which fall short due to the expensive data collection process and lack of capturing human behavior. In this paper, aiming at creating a cultural map from the user behavior perspective, we propose a participatory cultural mapping approach based on the collective behavior in LBSNs. Specifically, we first collect user participatory sensed behavioral data from LBSNs and then filter out noisy data from non-local users. Afterwards, by extracting the three key features, i.e., daily activity pattern, inter-city mobility and linguistic feature, we propose a cultural clustering method based on spectral clustering techniques. Finally, we generate a cultural map by visualizing these cultural clusters

on the map. Based on a large-scale user check-in dataset collected from Foursquare, we conduct both qualitative and quantitative evaluation of the proposed approach. The results show that our approach can subtly capture cultural information from user behavioral data in LBSNs, and create representative cultural maps. Comparing our cultural maps with those created by traditional cultural mapping approaches based on survey data, we observe not only important cultural correlations between them, but also interesting differences caused by some unique cultural features extracted from user behavioral data.

In the future, we plan to broaden this work in several directions. First, since different parts of a city may include diverse cultures (e.g., China town and Wall street in New York), we plan to explore cultural maps with a different geographical granularity, such as different districts in a city. Second, in order to augment user behavioral data in cultural mapping, we would like to capture user behavior in different LBSNs, such as Twitter and Facebook, etc. Finally, we plan to investigate the cultural regions with low social media penetration to further evaluate our framework.

REFERENCES

Yong-Yeol Ahn, Sebastian E Ahnert, James P Bagrow, and Albert-László Barabási. 2011. Flavor network and the principles of food pairing. *Scientific reports* 1 (2011).

LNF Ana and Anil K Jain. 2003. Robust data clustering. In *Proceedings of the 2003 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2003)*. 128–133.

Bernard M Bass. 1960. *Leadership, psychology, and organizational behavior.* Harper.

Sandro Bauer, Anastasios Noulas, Diarmuid O Séaghdha, Stephen Clark, and Cecilia Mascolo. 2012. Talking places: Modelling and analysing linguistic content in foursquare. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing*. 348–357.

James Bell. 2012. The World's Muslims: Unity and Diversity. (2012). http://www.pewforum.org/files/2012/08/the-worlds-muslims-full-report.pdf

Michael Harris Bond and Kwok Leung. 2009. Cultural Mapping of Beliefs About the World and Their Application to a Social Psychology Involving Culture. *Understanding culture: Theory, research, and application* (2009), 109.

Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Z Sui. 2011. Exploring Millions of Footprints in Location Sharing Services.. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM 2011)*. 81–88.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law distributions in empirical data. *SIAM review* 51, 4 (2009), 661–703.

Michael Cole and Jerome S Bruner. 1971. Cultural differences and inferences about psychological processes. *American Psychologist* 26, 10 (1971), 867.

Carole Counihan and Penny Van Esterik. 2012. *Food and culture: A reader*. Routledge.

Justin Cranshaw, Raz Schwartz, Jason I Hong, and Norman M Sadeh. 2012. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City.. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM 2012)*. 58–65.

Paul Du Gay and Michael Pryke. 2002. *Cultural economy: cultural analysis and commercial life*. Sage.

Graeme Evans and Jo Foord. 2008. Cultural mapping and sustainable communities: planning for the arts revisited. *Cultural trends* 17, 2 (2008), 65–96.

Scott A Golder and Michael W Macy. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* 333, 6051 (2011), 1878–1881.

Vipin Gupta, Paul J Hanges, and Peter Dorfman. 2002. Cultural clusters: Methodology and findings. *Journal of world business* 37, 1 (2002), 11–15.

Charles A Heatwole. 2006. Culture: A Geographical Perspective. (2006). http://www.p12.nysed.gov/ciai/socst/grade3/geograph.html

Peter Hemmersam, Jonny Aspen, Andrew Morrison, Idunn Sem, Martin Havnør, and Even Westvang. 2014. Exploring locative media for cultural mapping. *Mobility and Locative Media: Mobile Communication in Hybrid Spaces* (2014), 167.

Edward Adamson Hoebel. 1972. *Anthropology: The study of man*. McGraw-Hill New York.

Geert Hofstede. 1986. Cultural differences in teaching and learning. *International Journal of intercultural relations* 10, 3 (1986), 301–320.

Ronald Inglehart and Christian Welzel. 2010. Changing mass priorities: The link between modernization and democracy. *Perspectives on Politics* 8, 02 (2010), 551–567.

Claire Kramsch. 1998. *Language and culture*. Oxford University Press.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* (1951), 79–86.

Kevin N Laland, John Odling-Smee, and Sean Myles. 2010. How culture shaped the human genome: bringing genetics and the human sciences together. *Nature Reviews Genetics* 11, 2 (2010), 137–148.

Kalev Leetaru, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook. 2013. Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday* 18, 5 (2013).

Defu Lian, Xing Xie, Vincent W Zheng, Nicholas Jing Yuan, Fuzheng Zhang, and Enhong Chen. 2015. CEPR: A Collaborative Exploration and Periodically Returning Model for Location Prediction. *ACM Transactions on Intelligent Systems and Technology* 6, 1 (2015), 8.

Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* 37, 1 (1991), 145–151.

William C McGrew. 1998. Culture in Nonhuman primates? *Annual Review of Anthropology* 27, 1 (1998), 301–328.

Ann Mische. 2011. Relational Sociology, Culture, and Agency. In *The Sage Handbook of Social Network Analysis*, John Scott and Peter Carrington (Eds.). Sage, 80–97.

Robert T Moran, Philip R Harris, and Sarah Moran. 2007. *Managing cultural differences*. Routledge.

Andrew Y Ng, Michael I Jordan, Yair Weiss, and others. 2002. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems* 2 (2002), 849–856.

Anastasios Noulas, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil, and Cecilia Mascolo. 2012. A tale of many cities: universal patterns in human urban mobility. *PloS one* 7, 5 (2012), e37027.

Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. 2011. Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks. *The Social Mobile Web* 11 (2011).

Australia Department of Communications and the Arts. 1995. *Mapping culture: a guide for cultural and economic development in communities*. Canberra : A.G.P.S.

Jaram Park, Vladimir Barash, Clay Fink, and Meeyoung Cha. 2013. Emoticon style: Interpreting differences in emoticons across cultures. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM 2013)*. 466–475.

Charles Perreault and P Jeffrey Brantingham. 2011. Mobility-driven cultural transmission along the forager–collector continuum. *Journal of Anthropological Archaeology* 30, 1 (2011), 62–68.

Peter Poole. 2003. Cultural mapping and indigenous peoples. *A report for UNESCO* (2003).

Daniel Preoţiuc-Pietro, Justin Cranshaw, and Tae Yano. 2013. Exploring venue-based city-to-city similarity measures. In *Proceedings of the 2nd International Workshop on Urban Computing*. 16.

Roland Robertson. 1992. *Globalization: Social theory and global culture*. Vol. 16. Sage.

Edward Sapir. 1927. Language as a Form of Human Behavior. *The English Journal* 16, 6 (1927), 421–433.

Shalom H Schwartz. 2004. Mapping and interpreting cultural differences around the world. *International studies in sociology and social anthropology* (2004), 43–73.

Nakatani Shuyo. 2010. Language Detection Library for Java. (2010). http://code.google.com/p/language-detection/

Thiago H Silva, Pedro OS Vaz De Melo, Jussara M Almeida, and Antonio AF Loureiro. 2014. Large-scale study of city dynamics and urban social behavior using participatory sensing. *IEEE Wireless Communications* 21, 1 (2014), 42–51.

Burrhus Frederic Skinner. 1953. *Science and human behavior*. Simon and Schuster.

Walter W. Taylor. 1967. *A study of archeology*. Southern Illinois University Press.

Harry C Triandis. 1989. The self and social behavior in differing cultural contexts. *Psychological review* 96, 3 (1989), 506.

Ulrike Von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and computing* 17, 4 (2007), 395–416.

Zhu Wang, Daqing Zhang, Dingqi Yang, Zhiyong Yu, Xingshe Zhou, and Zhiwen Yu. 2012. Investigating city characteristics based on community profiling in LBSNs. In *Proceedings of the 2012 International Conference on Cloud and Green Computing*. 578–585.

Zhu Wang, Daqing Zhang, Xingshe Zhou, Dingqi Yang, Zhiyong Yu, and Zhiwen Yu. 2014. Discovering and Profiling Overlapping Communities in Location-Based Social Networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 44, 4 (April 2014), 499–509.

Dingqi Yang, Daqing Zhang, Longbiao Chen, and Bingqing Qu. 2015. NationTelescope: Monitoring and visualizing large-scale collective behavior in LBSNs. *Journal of Network and Computer Applications* 55 (2015), 170–180.

Dingqi Yang, Daqing Zhang, Zhiyong Yu, and Zhu Wang. 2013a. A sentiment-enhanced personalized location recommendation system. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media (HT 2013)*. 119–128.

Dingqi Yang, Daqing Zhang, Zhiyong Yu, and Zhiwen Yu. 2013b. Fine-grained preference-aware location search leveraging crowdsourced digital footprints from LBSNs. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing (UbiComp 2013)*. 479–488.

Dingqi Yang, Daqing Zhang, Zhiyong Yu, Zhiwen Yu, and Djamal Zeghlache. 2014. SESAME: Mining User Digital Footprints for Fine-Grained Preference-Aware Social Media Search. *ACM Transactions on Internet Technology* 14, 4 (2014), 28.

Dingqi Yang, Daqing Zhang, Vincent W Zheng, and Zhiyong Yu. 2015. Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 45, 1 (2015), 129–142.

Zhiyong Yu, Daqing Zhang, Zhiwen Yu, and Dingqi Yang. 2015. Participant Selection for Offline Event Marketing Leveraging Location-Based Social Networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 45, 6 (June 2015), 853–864.

Nicholas Jing Yuan, Fuzheng Zhang, Defu Lian, Kai Zheng, Siyu Yu, and Xing Xie. 2013. We know how you live: exploring the spectrum of urban lifestyles. In *Proceedings of the ACM Conference on Online Social Networks (COSN 2013)*. 3–14.

Yu-Xiao Zhu, Junming Huang, Zi-Ke Zhang, Qian-Ming Zhang, Tao Zhou, and Yong-Yeol Ahn. 2013. Geography and similarity of regional cuisines in China. *PloS one* 8, 11 (2013), e79161.