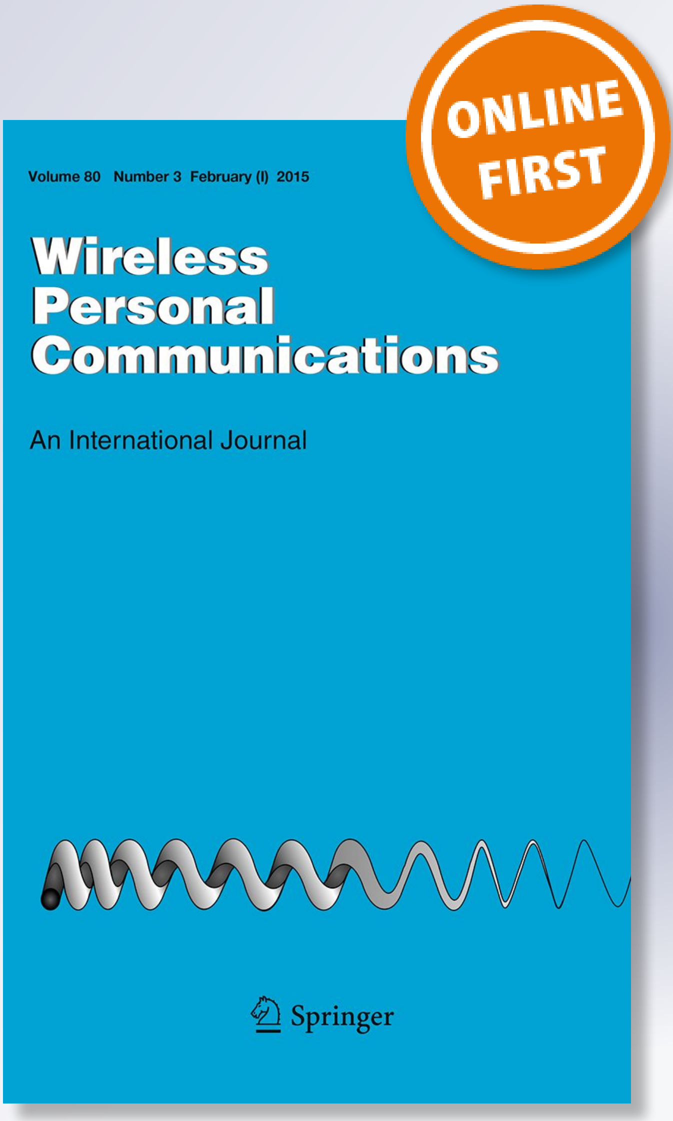
Measuring Geospatial Properties: Relating Online Content Browsing Behaviors to Users' Points of Interest

Qiujian Lv, Yuanyuan Qiao, Yi Zhang, Fehmi Ben Abdesslem, Wenhui Lin & Jie Yang

Wireless Personal Communications
An International Journal

ISSN 0929-6212

Wireless Pers Commun
DOI 10.1007/s11277-018-5773-7



ONLINE FIRST

Volume 80 Number 3 February (I) 2015

Wireless Personal Communications

An International Journal

Springer

Springer

Wireless Pers Commun
https://doi.org/10.1007/s11277-018-5773-7

Measuring Geospatial Properties: Relating Online Content Browsing Behaviors to Users' Points of Interest

Qiujian Lv[1] · Yuanyuan Qiao[2] · Yi Zhang[2] · Fehmi Ben Abdesslem[3] · Wenhui Lin[4] · Jie Yang[2]

Abstract With the growth of the Mobile Internet, people have become active in both the online and offline worlds. Investigating the relationships between users' online and offline behaviors is critical for personalization and content caching, as well as improving urban planning. Although some studies have measured the spatial properties of online social relationships, there have been few in-depth investigations of the relationships between users' online content browsing behaviors and their real-life locations. This paper provides the first insight into the geospatial properties of online content browsing behaviors from the perspectives of both geographical regions and individual users. We first analyze the online browsing patterns across geographical regions. Then, a multilayer-network-based model is presented to discover how inter-user distances affect the distributions of users with similar online browsing interests. Drawing upon results from a comprehensive study of users of

✉ Qiujian Lv
lvqiujian@bupt.edu.cn

Yuanyuan Qiao
yyqiao@bupt.edu.cn

Yi Zhang
yi_zhang@bupt.edu.cn

Fehmi Ben Abdesslem
fehmi@sics.se

Wenhui Lin
linwenhui@aisino.com

Jie Yang
janeyang@bupt.edu.cn

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

[2] Research Center of Network Monitoring and Analysis, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China

[3] Decisions, Networks and Analytics Laboratory, Swedish Institute of Computer Science, Stockholm, Sweden

[4] Technology Research Institute, Aisino Corporation, Beijing, China

Published online: 07 May 2018

three popular online content services in a metropolitan city in China, we achieve a broad understanding of the general and specific geospatial properties of users' various preferences. Specifically, users with similar online browsing interests exhibit, to a large extent, strong geographic correlations, and different services exhibit distinct geospatial properties in terms of their usage patterns. The results of this work can potentially be exploited to improve a vast number of applications.

Keywords Online content services · Online content browsing · Geospatial properties · Multilayer network

1 Introduction

Between the debut of the second-generation (2G) mobile communication system in 1991 and the launch of the 4G system in 2010, mobile networks have transformed from a pure telephony system into a true Mobile Internet that can transmit rich multimedia content such as online games, e-books, e-music, and videos [44]. Many aspects of our daily lives have been permeated by various online content services. In addition to participating in offline activities, individuals are increasingly enjoying content browsing activities on these online content services such as shopping on *Amazon*, buying vouchers on *Groupon*, watching movies on *YouTube*, and interacting with friends on *Facebook*. Various online content services have been analyzed by many researchers to gain an understanding of users' online preferences [24, 34, 46, 49] and online content purchasing intentions [15, 27, 45].

Thanks to new location-sensing technologies, location has become a crucial aspect related to many online services and is increasingly becoming a critical factor in bridging the gap between the online and offline worlds [51]. Tobler's first law of geography [42] states that "everything is related to everything else, but near things are more related than distant things". However, the question of whether a user's online behavior is more closely related to that of nearby users than to that of users who are farther away has received only limited investigation. The literature has mainly explored the relationship between geographical distances and online friendships [17, 38, 39]. The geospatial properties of users' browsing behaviors on online content service platforms have not yet been studied in detail. In addition, among all the locations of a user, some are visited less often or only sporadically, while other locations are points of interest (POIs). These POIs are associated with the semantics of a human's latent states such as their home or workplace. Based on the stay duration, a user's POIs can be distinguished into first POI, second POI, etc. Various studies have found that a user spends, on average, more than 50% of an entire observation period at his or her **f**irst **P**OI (*fp*) [2, 28]. Relating content browsing behaviors to users' locations of greatest interest, namely, their first POIs, can not only improve the understanding of general geospatial properties but also support location-based services for practical applications.

Investigating the geospatial properties of users' browsing behaviors on online content services, such as how users' locations are related to their behaviors regarding online shopping or movie watching, is critical for improving both online and offline services to make our daily lives easier [51]. Urban planning in the offline world is guided by an understanding of the specific online preferences of users in different regions to facilitate the daily lives of individuals. It also yields insight for online service providers with regard to the generation of personalized recommendations, especially for overcoming the cold-start problem [48]. The cold-start problem always refers to the fact that recommender systems tend to fail when little history about the user online content browsing behavior is known. In addition, the discovery of potential new friends who share both online browsing similarities and geographical correlations can be enabled by the comprehensive profiling of users in terms of both online and offline interests [17]. This ability to connect with more like-minded friends will enrich users' online and offline lives. In addition to online items of common interest, mutual recommendations of offline activities can be offered to expand their circles of offline friends.

To comprehensively determine the geospatial properties of users' browsing behaviors on online content services, it is important to simultaneously consider multiple popular online content services. Based on the application category dictionary presented in [47], online content services can be categorized into 10 groups. In addition to the *social network* category, we select five additional categories that represent interests and goals that can serve as triggers for users when arranging offline activities: *e-commerce*, *reading*, *video*, *music*, and *online gaming*. Among them, *e-commerce* dominates in terms of the number of users [47], whereas *video* traffic on mobile devices currently accounts for 46% of all IP traffic and is expected to continue growing [9]. Considering the popularity of these two categories, we focus on services related to *e-commerce* and *video* to study their geospatial properties. Moreover, we consider a specific popular e-commerce mode known as *group buying* [8], which provides the novel ability to "purchase online, redeem offline" [26] and is offered through location-based services (LBSs). Such a service identifies a user's location at a given time [37] and provides coupons for shops in the user's vicinity. *Group buying* is currently attracting a large number of users [8]. Intuitively, its novel features may result in distinct geospatial properties of *group buying* compared with common *e-commerce* services, and this intuition inspires us to analyze these two types of services separately. In light of these considerations, this paper focuses on three popular online content services—one for *group buying*, one for common *e-commerce*, and one for *video*—and relates their usage to the physical locations of their users to evaluate their geospatial properties. In this way, we ensure that our observations will not be specific to a particular type of service.

Based on multiple popular online content services, this paper presents an in-depth investigation of the geospatial properties of users' content browsing behaviors for a large-scale user population in a metropolitan city in China from the perspectives of both geographical regions and individual users. From the perspective of regions, we determine *which regions exhibit similar online browsing patterns and how regional features affect a region's online browsing behaviors*. Furthermore, at the more fine-grained level of individual users, we concentrate on both users in the same region and users distributed across various regions of the city. We determine *How are users who share similar online interests correlated with other users located in the same region?* and *What are the spatial distributions of users who share similar online browsing interests in a metropolitan city?* These quantitative results obtained for different online content services from the two perspectives reveal new insights regarding location-aware recommendation systems. Overall, the contributions of this paper can be summarized as follows:

1. To the best of our knowledge, we are the first to study the geospatial properties of users' online content browsing behaviors from the perspectives of both geographical regions and individual users. An extensive analysis is conducted based on various services, the data for which are extracted from real cellular network traffic data collected from a metropolitan city that represent the activities of more than 100,000 people in 3 months.

2. We cluster regions with similar online browsing behavior patterns and discover the pattern similarities between adjacent regions. Distinct phenomena are observed for different services. Moreover, the browsing patterns of regions are inferred to be related to certain regional features.
3. Using a multilayer network model, we present a thorough analysis of the geospatial properties of the usage of online content services at the level of individual users. We explore how the locations of users who share similar online interests are correlated. This in-depth quantification provides insight into the distinct correlations observed for different services. Moreover, we highlight the necessity of location awareness for recommendation systems.

The remainder of this paper is organized as follows. In Sect. 2, related works are reviewed. Sect. 3 describes the preliminaries for this paper, including the data collection methods and data characteristics, as well as the methods used to extract users' locations and online browsing behaviors. Sections 4 and 5 present the geospatial analyses at the levels of regions and individuals, respectively. Finally, conclusions are presented in Sect. 6.

2 Related Work

We categorize existing works related to the geospatial properties of users' content browsing behaviors into two types: (a) analyses of online content services and (b) analyses of the spatial properties of user behaviors.

2.1 Analyses of Online Content Services

Extensive research has been directed toward gaining separate understandings of different types of services, especially the two most popular service types: *e-commerce* [8, 24, 27] and *video* [5, 19, 43]. Since *e-commerce* is becoming a primary means for consumers to find, compare, and ultimately purchase products [27], consumer preferences [24, 34, 46, 49] and online purchasing intentions have been investigated by many researchers [15, 27, 45]. Regarding *video* services, the properties of videos and how their popularity can be modeled, classified and predicted [7, 11, 14] have been measured in some studies, although other studies have modeled the quality of experience (QoE) as perceived by users when watching movies [19, 43]. Recently, instead of studying the services themselves, some researchers have analyzed the impact of geolocation on online services. They have found that e-commerce follows residential mobility patterns [6, 35] and affects the number of shopping trips made by users [10]. These researchers have also encouraged the development of spatio-temporal methods of acquiring fine-grained localization information [23] with the emergence of *group buying*. However, in these studies, the personal information of customers has been collected through surveys, which are time intensive, expansive, and necessarily limited to small groups of users. Concerning *video* services, researchers have demonstrated that online video consumption appears to be constrained by geographic locality [5]. The analysis addressed the view counts of videos across regions; however, the geographic popularity of videos was not analyzed at the more fine-grained level of individual users. Moreover, all the studies discussed above treated different types of services separately and failed to find any general or specific properties of diverse online content services, especially geospatial properties related to users' various online preferences.

2.2 Spatial Properties of User Behaviors

The proliferation of Mobile Internet access has broadly affected people's way of life [6, 10, 35], and users are currently living in an era of mixed online and offline activities [53]. The effect of spatial location on online user behavior has recently been addressed [5]. In the literature, datasets collected from location-based social networks (LBSNs) have mainly been used to analyze the spatial properties of online social relationships [12, 13, 17, 38, 39]. By analyzing *Facebook*, *Twitter*, and face-to-face networks, the structural characteristics of online social communities and offline face-to-face networks being similar was discovered by Dunbar et al. [13]. Social relationships on *Facebook*, phone communications, and user profiles for 74 students were collected by Hristova et al. [17], who used these data to investigate homophily among multiplex social ties. Scellato et al. [38] studied four online social service networks along with geographic information and defined two metrics: the *node locality* and a *geographic clustering coefficient*. All the works discussed above leveraged explicit social ties between users and applied graph-analysis-based approaches to analyze the spatial properties of online social relationships. By contrast, we consider implicit relationships between users based on similar online behaviors or visited locations [20]. Then, we adopt a graph-analysis-based method to discover the geospatial properties of users' online content browsing behaviors.

Moreover, increasing numbers of studies have begun to investigate the correlations between user online behaviors and regions of a city. They are motivated by the potential existence of close relationships between online and offline behaviors, and they present data-driven methods for urban planning [36]. Some studies have associated users' online behaviors with the regions where the online behaviors are generated, and they have conducted geospatial analysis in terms of bytes, packets, flow counts [33, 40], and Application (app) usage [40]. For example, *k*-means was applied by Shafiq et al. [40] to cluster cells with similar app usage patterns; they found that cell locations corresponding to the usage of particular applications tended to be co-located. Meanwhile, a series of works have aimed to study the geographic distribution of a topic in terms of user-generated social media [21, 50] using a probabilistic topic model. Latent Dirichlet Allocation (LDA), a baseline and powerful topic model, provides a natural way to enhance the interpretability of such analysis for decision making. LDA was applied by Kling [21] to obtain a decomposition of the stream of digital traces of citizens in a set of city-scale activities called urban topics. The space-time structures of the topical content of short textual messages from Twitter was explored by Pozdnoukhov [36] using a streaming LDA topic model. However, the above works failed to study the correlations between online content browsing behaviors and regions. By contrast, this paper applies LDA to the obtained regional online browsing pattern, based on which the geospatial properties of users' online browsing behaviors from the perspective of regions are revealed.

3 Preliminaries

Here, we introduce our data collection methodology and describe the collected data. Then, we explain how we extract users' locations and online browsing behaviors.

3.1 Data Collection

Recently, with the growing prevalence of cellular networks, users have become able to access online content via their mobile devices while in transit from one location to another. This has the potential to leave traces in the networks of users' online behaviors and offline mobility patterns. To analyze these behaviors, we have developed a Traffic Monitoring System (TMS) [25] that can collect the data generated by users in a cellular network. As shown in Fig. 1, instances of our TMS are deployed at multiple interfaces with the core network of a major cellular operator such as the interface between a Serving GPRS Support Node (SGSN) and a Gateway GPRS Support Node (GGSN). They monitor packets, aggregate packets into flows in real time, and produce sequences of time-stamped records, each of which contains a user-anonymized identifier (ID), the cell ID, and the accessed Uniform Resource Locator (URL), etc. [25]. The services being used are identified based on keywords found in the URLs such as *maps.google* or *Amazon*. The longitude and latitude associated with the cell ID provide the user's location, and the URL indicates the content browsed by the user.

3.2 Data Description

The dataset contains records from over 182,916 mobile phone users and 10,809 cells in an urban area. The dataset was collected from October 1st, 2015, to December 31st, 2015, and consists of 6,211,434,723 records. To improve the quality of the analysis, the experiments presented below focus on active users who accessed the cellular data network on more than 60 different days. In addition, to obtain comprehensive geospatial properties of users' browsing behaviors when utilizing online content services, we focus on three popular online content services in China: a common *e-commerce* service, a *group buying* service, and a *video* service.

(a) ***Suning*** (common *e-commerce* service): In China, *Taobao* is the most visited online shopping service. However, *Taobao* uses Hyper Text Transfer Protocol over Secure

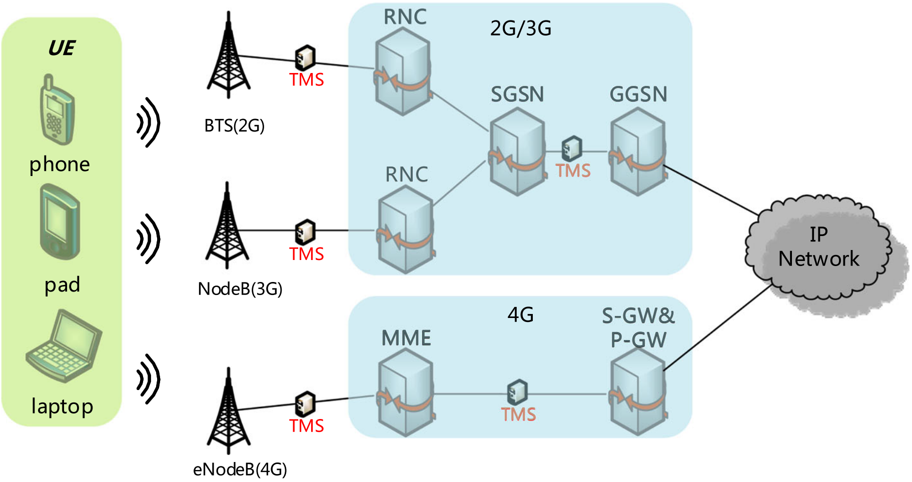


Fig. 1 A cellular data network instrumented with TMS data capture devices

Sockets Layer (HTTPS) as its transmission protocol, which makes it impossible to obtain detailed information on user content browsing behaviors. Therefore, we consider another popular *e-commerce* service, ***Suning***, which ranks among China's top three business-to-consumer (B2C) companies. The product categories in which ***Suning*** operates are computer, communication, and consumer electronics (3C) products; books; household commodities; and cosmetics. We extracted data on 4086 active ***Suning*** users for analysis.

(b) ***Meituan*** (*group buying* service): ***Meituan*** is one of China's earliest and most successful group buying services, similar to *Groupon*. It operates as an LBS, and the items that it presents are shops offering discounts for users in the city. Items of this type are associated with explicit location information. We extracted data on 33,647 active ***Meituan*** users for analysis.

(c) ***Youku*** (*Video* service): ***Youku*** is a leading Chinese online video service platform similar to *YouTube*. The items presented by ***Youku*** are videos. We extracted data on 24,729 active ***Youku*** users for analysis.

These three services are all popular online content services in China. The nature of ***Meituan*** as an LBS is essential to the services it offers, whereas ***Suning*** and ***Youku*** place less emphasis on location. Based on the high-volume records collected from these services, this paper presents the first investigation of the general and specific geospatial properties of users' browsing behaviors on various online service platforms.

3.3 Extraction of Users' Locations

Spatial trajectories can be unintentionally generated when an individual moves from one cell to another while accessing the Mobile Internet. These trajectories are represented by sequences of cell IDs with their corresponding transition times. This section presents the extraction of Points of Interest (POIs) from user trajectories, upon which users' locations are obtained.

In our previous study [28], we identified each user's POIs from their trajectory using a "Leader-Follower Clustering"-based method. This technique depends on two parameters: a radial distance threshold, D_r, and a time frequency threshold, T_r. To select suitable thresholds, the technique discovered the turning points of the curves, which indicate how the average number of POIs changes as a function of the threshold. The turning points signify the thresholds immediately before the number of cell IDs begins to converge to the number of POIs [28]. Thus, POIs are identified as areas with a radius of 1000 meters that are visited on more than 31.8% of the observed days.

Then, we concentrate on the location of greatest interest *fp* and divide the latitude and longitude dimensions of all users' first POIs into discrete 0.01×0.01 latitude/longitude regions (approximately $1000\,\mathrm{m} \times 1000\,\mathrm{m}$) to locate users in a global 2-dimensional metric space. Each region is represented by the latitude and longitude of its centroid. Figure 2 illustrates the method used to discretize the users' first POIs for an example consisting of 2 regions.

Finally, a user's location is defined as the region in which their *fp* is located. Later in the paper, we will relate users' online content browsing behaviors to their first POIs and show how we can infer correlations between users' online browsing behaviors and locations at the levels of regions (Sect. 4) and individuals (Sect. 5).

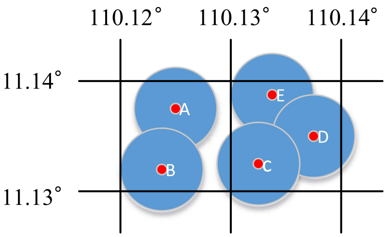


Fig. 2 Discretizing users' first POIs into 2 regions. The black rectangles delineate the regions. A, B, C, D, and E represent five distinct users. Each blue circle represents the area of the corresponding user's first POI. The red dot in each circle denotes its centroid

3.4 Extraction of Users' Online Browsing Behaviors

The services that we study offer a large number of possible items that users can access, although users will click on only a few items during a given time period [1]. Therefore, it is infeasible to study inter-user browsing behavior correlations on a per-item basis; instead, it is necessary to group items into more general categories. Luckily, to facilitate the effective management of a large number of items, each service defines a set of categorization tags $Tags = \{tag_1, tag_2, \ldots, tag_n\}$ and assigns a selection of tags $Tags_{item_i}$ to each item $item_i$, where $Tags_{item_i} \subset Tags$. Taking ***Meituan*** and ***Suning*** as examples, Fig. 3 illustrates the assignment of tags to items. The tag set $Tags_{item_i}$ represents the features of $item_i$. Therefore, the tags associated with items that users have browsed can reveal those users' interests. The online browsing behavior of user u_i can thus be represented by the aggregated tags of n items that they have browsed, i.e.,

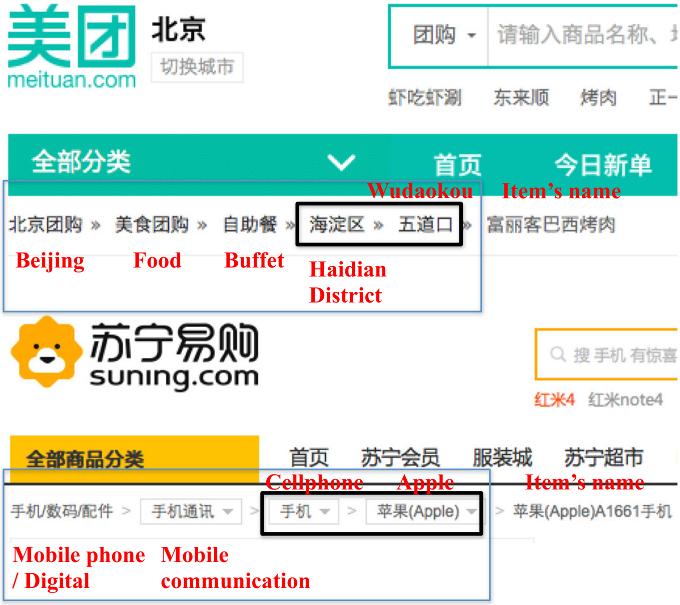


Fig. 3 Examples of categorization tags for items in ***Meituan*** and ***Suning***. The large boxes contain all tags for an item, whereas the small boxes indicate lower level (more specific) tags

$$UB_i = \{Tags_{item_1}, Tags_{item_2}, \ldots, Tags_{item_n}\}$$

The item ID, short for the distinct identifier of an item in a service, can be extracted from the related URL through the matching of regular expressions [22]. By sending an item ID in a formatted Hyper Text Transfer Protocol (HTTP) GET request to the service's Application Programming Interface (API), the item's tags can be retrieved as a JavaScript Object Notation (JSON) object. Using this method, the lower level tags, which are most relevant to an item's features, are collected: (a) the tags collected from ***Suning*** represent the brands or categories of items, such as {*Apple* or *cellphone*}; (b) the tags collected from ***Meituan*** mainly represent the business centers where the items are located, such as {*Wudaokou* or *Haidian District*}; and (c) the tags of items in ***Youku*** describe their genres such as {*comedy* or *American*}.

Different services manage different numbers of tags, and the different tags of a given service are also browsed by different numbers of users. Thus, it is essential to select meaningful tags from the three services to characterize users' online browsing behaviors. The details of tag selection are as follows:

(1) Certain tags, such as *TV series* and *romantic play* on ***Youku***, are browsed by users in more than 90% of the regions of the city. To reduce the influence of such popular tags on the ability to discover distinct behaviors among users distributed across the city, these tags are filtered out in advance.

(2) We rank the tags according to the proportion of users who have browsed items with those tags. Many tags are browsed by only a small fraction of users, as shown in Fig. 4a for ***Meituan***. Therefore, a suitable threshold is needed for filtering out tags that are browsed by relatively few users. To find the optimal threshold, a knee of (110, 0.25%) is found in the curve of ***Meituan***, where a significant change in the slope of the curve is observed [28]. The knee separates the meaningful tags from other tags browsed by few users [28]. In addition, the rank distributions of ***Youku*** and ***Suning*** reach their knees at higher ratios of users (considering space limitations, the rank distributions for ***Youku*** and ***Suning*** are not shown). Hence, to obtain all the meaningful tags of the three services, a threshold of 0.25% is chosen, and tags browsed by less than 0.25% of users are filtered out.

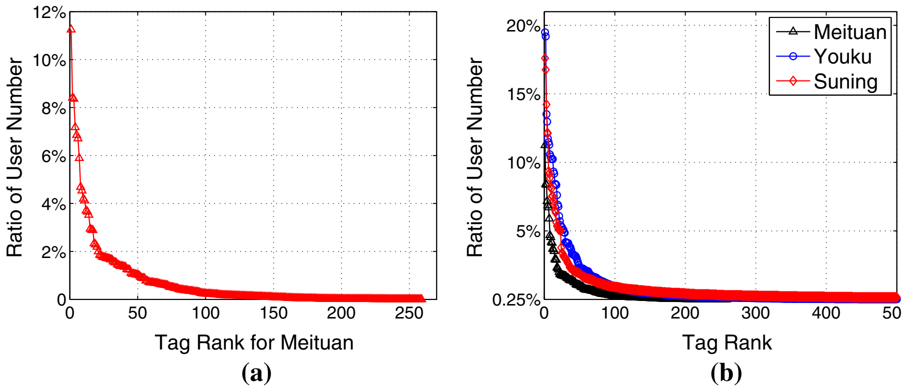


Fig. 4 **a** Rank distribution of tags in ***Meituan***. **b** Rank distributions of the tags remaining after filtering for all three services

Subsequently, the numbers of remaining tags for ***Meituan***, ***Youku***, and ***Suning*** are 110, 195, and 303, respectively. The original total numbers of tags for these three services are 257, 564, and 2791, respectively. The rank distributions for the remaining tags are depicted in Fig. 4b, which shows that the three services exhibit similar distributions.

Notably, although the semantics implied by the tags used by the different services are distinct, for our subsequent comparative analyses, it is appropriate for us to treat the tags from different services equally in characterizing users' online browsing behaviors for the following reasons. First, after filtering, the three services have similar quantities of tags, and the rank distributions of the remaining tags show similar features. Second, all extracted tags are drawn from the lower level tags used to characterize items. Thus, they offer similar functionalities for their corresponding services. They reflect users' habits in terms of item selection and allow users to effectively search for items. For instance, ***Meituan*** users are mostly concerned with where they are or where they will go [26], whereas ***Suning*** users care about the brands or categories of the items that they need, and they choose their items of interest accordingly. Furthermore, service providers always offer recommendations based on these tags [24].

4 Geospatial Analysis at the Region Level

One of the main concerns of this paper is to determine *which regions exhibit similar online browsing patterns and how regional features affect a region's online browsing behaviors*. Such an investigation will reveal the particular online preferences of users in different regions and provide new insights into location-aware targeted marketing. To this end, this section follows a three-step methodology to address this issue through a region-level analysis. First, by considering users whose first POIs are in the same region, we derive regional online browsing behaviors. Second, we determine regional online browsing patterns using an LDA model. Third, the identified regional online browsing patterns are grouped using a clustering algorithm.

4.1 Obtaining Regional Online Browsing Behaviors

As the first step, we extract users whose first POIs are in the same region, called *co-poi* users. For the example shown in Fig. 2, the *co-poi* user groups are {A, B} and {C, D, E}. Based on those *co-poi* users, we then determine the regional online browsing behaviors, which represent the general browsing behaviors of all users in a region and are derived as follows:

(a) Obtaining regional behavior vectors. The regional behavior vector RB_i represents the online browsing behavior of users in region r_i. Similar to UB_i, RB_i is composed of the tags of items browsed by users whose first POIs are located in region r_i. To obtain RB_i, we first remove duplicate tags from the online browsing behavior UB_i of each user u_i to obtain a behavior vector $UB_i^{'} = (tag_1, tag_2, \ldots, tag_n)$. Duplicate removal helps to decrease the influence of specific users' browsing behaviors on the behavior pattern of a region. Then, the regional behavior vector RB_i is formed by aggregating the behavior vectors of all n users whose first POIs (*fp*) lie within region r_i, i.e., $RB_i = (UB_{1i}^{'}, UB_{2i}^{'}, \ldots, UB_{ni}^{'})$.

(b) Deriving regional tag vectors. The regional tag vector f_i represents the importance of the tags in the regional behavior vector RB_i of region r_i. To adjust for the fact that some tags appear more frequently in general, we calculate the term frequency-inverse document frequency (TF-IDF) values [52]. We then construct $f_i = (v_{i1}, v_{i2}, \ldots, v_{in})$, where v_{ij} is the TF-IDF value of the j-th tag and n is the number of tags in RB_i. The TF-IDF value v_{ij} is given by

$$v_{ij} = \frac{n_j}{N_i} \times \log \frac{R}{||\{RB_i | tag_j \in RB_i\}||} \tag{1}$$

where n_j is the number of instances of tag_j in RB_i, N_i is the number of tags in RB_i, R is the number of regions, and $||\{RB_i | tag_j \in RB_i\}||$ is the number of regions that contain tag_j.

Finally, the regional tag vectors are obtained to represent regional online browsing behaviors, based on which we can derive regional online browsing patterns.

4.2 Deriving Regional Online Browsing Patterns

Although the tags that represent regional online browsing behaviors are diverse, some of them are semantically related and represent common online interests. To correctly reveal regional online preferences, semantically related tags need to be summarized to define more general online browsing activities. In this way, the online browsing patterns of each region, represented by distributions of online browsing activities, can be derived. To this end, this section first describes the derivation of regional online browsing patterns based on an LDA model and then presents the formulation of an optimal LDA model. Finally, the online browsing activities identified using this optimal model are shown.

LDA [3], a popular and powerful topic model, was developed to model document content based on the idea that the probability distribution over words in a document can be expressed as a mixture of topics, where each topic is a probability distribution over words. The nature of LDA is concordant with the problem of obtaining the online browsing patterns of each region by summarizing semantically related tags. Thus, an LDA-model-based solution is proposed by establishing analogies between the task of identifying the latent online browsing activities in a region and the problem of discovering the latent topics of a document. As shown in Table 1, we treat the tags as words and the regional tag vectors as the metadata of documents. We regard a region as a document and a browsing activity as a topic. In the process of deriving regional online browsing patterns, the LDA model first produces φ_k^{tag} and $\theta_{r_i}^k$, which represent the probability of a tag being associated with browsing activity k and the probability of a browsing activity k being associated with region r_i, respectively. Given these probability distributions, we rank the tags for each discovered browsing activity and determine the browsing pattern for each region. For region r_i, the browsing pattern is a K-dimensional vector $\theta_{r_i} = (\theta_{r_i}^1, \theta_{r_i}^2, \ldots, \theta_{r_i}^K)$, where K is the number of browsing activities and $\theta_{r_i}^k$ is the proportion of activity k. Considering space

Table 1 Analogies from regional browsing activities to document topics

Tags	→	Words
Regions	→	Documents
Regional tag vector of a region	→	Metadata of a document
Browsing activities of a region	→	Topics of a document

limitations, the details of the implementation of the popular LDA model are not described here and can be found in [3]. Rather than directly using a clustering method, this method helps to group words into semantic topics, based on which the obtained regional online browsing patterns are better interpreted [52].

The number of latent topics (or browsing activities) K is an important parameter in determining an optimal LDA model. Seeking too many topics may lead to over-fitting and make the learned results difficult to interpret. The *perplexity* [3] is a measure used to determine the optimal number of latent topics K. For K values ranging from 0 to 100 in increments of 10, we compute the *perplexity* for each LDA by means of 1000 iterations of the Gibbs sampling algorithm [3]. For ***Youku*** and ***Meituan***, a drop in *perplexity* occurs for $K < 30$, after which the *perplexity* stabilizes. Thus, we choose $K = 30$ for the construction of our LDA model. Using the same method, the number of latent browsing activities for ***Suning*** is set to 10.

When identifying relevant online browsing activities (topics) using the optimal LDA model, several tags are assigned with a certain probability. To characterize these activities, we rank the tags by φ_k^{tag} for each topic k and list the top-ranked tags accounting for more than 90% of the total probability. Several interesting results are found as follows:

(a) ***Meituan*** For visualization, the categorization tags are replaced with the latitude and longitude of each business center, with two digits after the decimal point. Figure 5 plots the results for several topics on a map. Each circle with a red dot in the center represents a tag covering a 0.01 × 0.01 latitude/longitude region. Each topic is shown to be composed of geographically concentrated regions and spans several adjacent business centers.

(b) ***Suning*** Table 2 presents the results for several topics related to ***Suning***. Each topic represents a distinct online shopping interest. User interest in cellphones is captured by Topic 4, for which all related tags are mobile phone brands such as Vivo, Nokia, HTC, and iPhone. By contrast, the tags of Topic 9 are all related to household appliances such as washing machines, water heaters, and refrigerators (Whirlpool and Little Swan are household appliance brands).

(c) ***Youku*** Table 3 presents the results for several topics identified in relation to ***Youku***. Each topic contains several tags with similar semantic meanings and represents a particular type of online video watching activity. For instance, Topic 11 represents video browsing activity related to comedy, whereas Topic 1 reflects interest in gangster videos produced in Hong Kong.

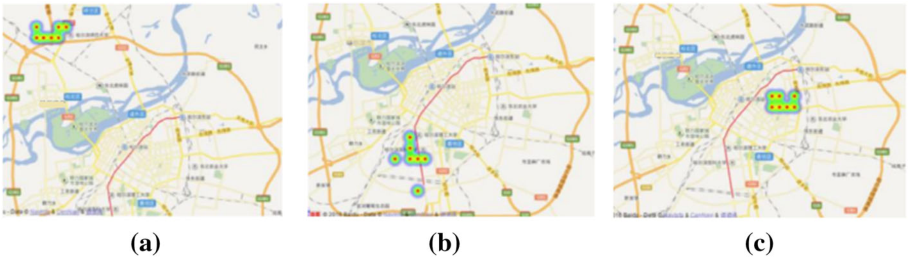


Fig. 5 ***Meituan***: Visualizations of topics based on their most likely associated words (tags). **a** Topic 5, **b** Topic 8, **c** Topic 13

Table 2 ***Suning***: characterization of topics in terms of their most likely associated words (tags)

Topic 1	Tag	Daily necessity	Hair care	Shampoo	Storage article
	φ_k^{tag}	0.04	0.03	0.03	0.02
Topic 4	tag	Vivo	Nokia	Smartisan	HTC
	φ_k^{tag}	0.06	0.03	0.03	0.02
Topic 5	Tag	Cream	Face cleaning	Shower gel	Body bath
	φ_k^{tag}	0.02	0.02	0.02	0.02
Topic 9	Tag	Washing machine	Whirlpool	Little Swan	Water heater
	φ_k^{tag}	0.02	0.02	0.02	0.02

Table 3 ***Youku***: characterization of topics in terms of their most likely associated words (tags)

Topic 0	Tag	Anime list	Anime channel	Anime appreciation
	φ_k^{tag}	0.26	0.21	0.15
Topic 1	Tag	Hong Kong	Suspense	Crime
	φ_k^{tag}	0.9	0.07	0.01
Topic 3	Tag	Parent-child channel	Parent-child list	Baby show
	φ_k^{tag}	0.27	0.25	0.21
Topic 11	Tag	Funny list	Story play	Comedy
	φ_k^{tag}	0.50	0.36	0.04

In summary, although the tags used in the various services are diverse, the topics discovered via LDA consist of tags with similar meanings and are representative of very specific activities, despite the inherent noise present among various individuals in a region.

4.3 Aggregating Regions with Similar Browsing Patterns

To understand which regions exhibit similar online browsing patterns and how regional features affect a region's online browsing behaviors, we aggregate regions with similar activity (topic) distributions into *k* clusters using a clustering algorithm. We apply *k*-means clustering to the *K*-dimensional points θ_{r_i}, $i \in 1, 2, \ldots, R$. The number of clusters *k* is determined based on the average Calinski-Harabasz index (*CHindex*) [32]. In practice, we perform cross-validation multiple times for different *k* values and choose the value of *k* with the maximum overall *CHindex* value. As a result, we identify 20 clusters for ***Meituan***, whereas 3 clusters are formed for ***Suning*** and ***Youku***. The aggregated regions are visualized in Fig. 6.

Figure 6 indicates that users in adjacent regions are, in general, likely to be in the same cluster. They enjoy group buying services for the same business centers, watch videos with similar tags, or buy the same types of goods. Meanwhile, the distinct numbers and shapes of the groups for the three services reflect the distinct correlations between users' online browsing behaviors and their locations with regard to different types of services. The users of an LBS such as ***Meituan*** who share similar online browsing behaviors are clustered into confined regions. By contrast, for services that are less specialized based on user locations,

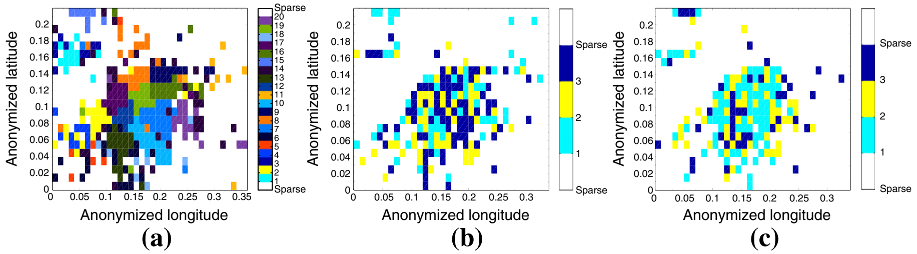


Fig. 6 Aggregated region groups with respect to the three services. Regions shown in the same color belong to the same group. Regions presented in white contain fewer than ten users and are ignored during clustering. **a** ***Meituan***, **b** ***Youku***, **c** ***Suning***

such as ***Suning*** and ***Youku***, regions sharing similar online browsing interests have a larger geographical spread.

Furthermore, we analyze the browsing patterns of the aggregated groups for each service. The differences between the results for ***Youku*** and ***Suning*** are very small, resulting in similar conclusions. Hence, considering space limitations, we present the detailed analysis results only for ***Suning*** and ***Meituan***.

(a) ***Meituan*** Figure 7 shows the geographic locations of several clusters based on the first POIs of the users in those clusters (left) and their corresponding activity patterns regarding online shop browsing (right). Each dot represents a 0.01 × 0.01 region of the latitude/longitude grid. A dot with a darker color in the graph on the right-hand side indicates that the topic represented by that dot is more likely to be browsed by users whose first POIs appear in the corresponding graph on the left-hand side. The *fp* locations are closest to the topics with the darkest colors. For topics corresponding to locations farther away from the users' first POIs, the shops related to these topics are browsed less frequently. Hence, for this LBS, the browsing activities of a user tend to be geographically confined to shops near their first POI.

(b) ***Suning*** The geolocation data of three clusters are visualized in Fig. 8, and their dominant topics are reported in the caption. Adjacent regions tend to belong to the same cluster, and different clusters exhibit distinct activity patterns. The browsing patterns of each region can be inferred to be correlated with the features of that region as follows:

Cluster 1: The predominant topics concern household appliances such as electric stoves and washing machines. The regions dominated by these topics cover several new residential blocks. Moving into new houses motivates residents to buy new household appliances.

Cluster 2: The regions corresponding to this cluster include several universities, a railway station and the old city area. The interests of the residents living here are focused on general merchandise such as household commodities, baby care products, and cosmetics.

Cluster 3: The topics associated with this cluster are related to 3C products such as smartphones, cameras, and air cleaners. These regions encompass the economic development zone and business districts of the city. Users who spend considerable amounts of time here are typically high-

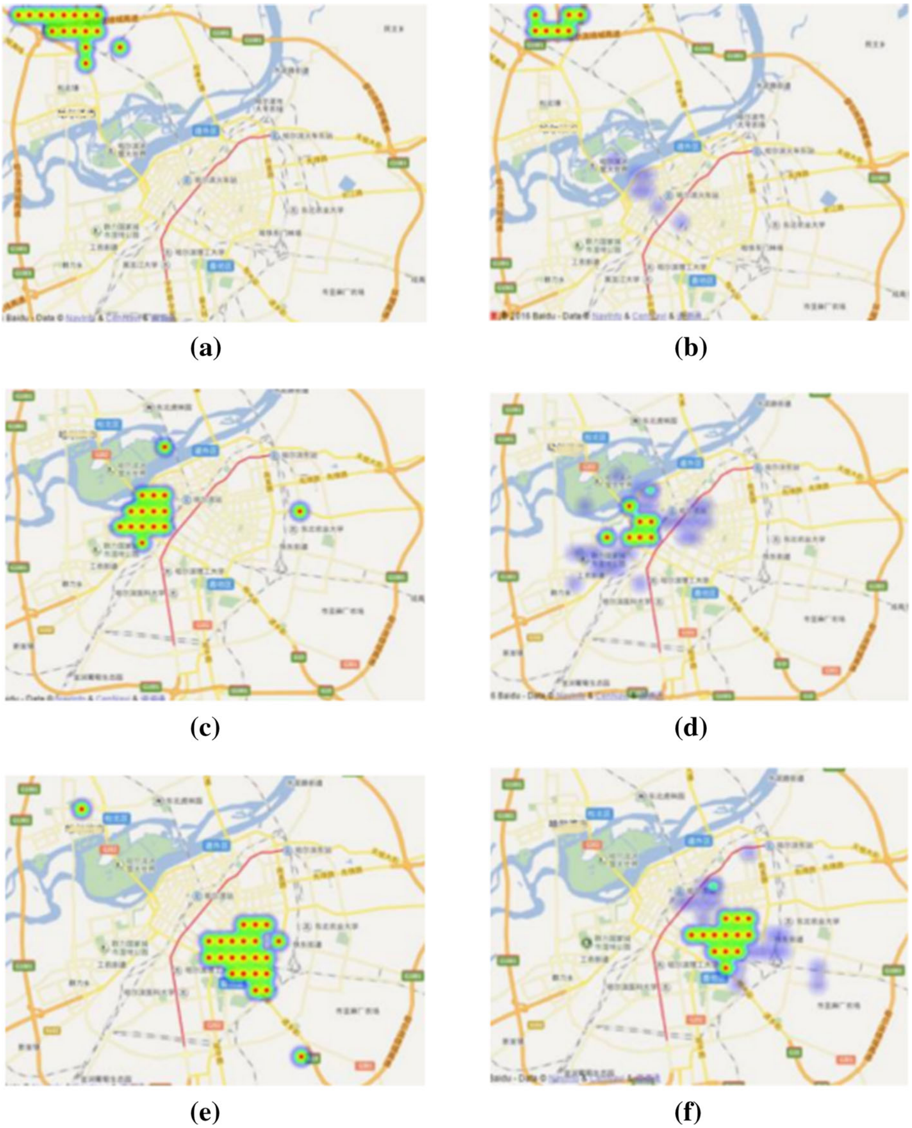


Fig. 7 ***Meituan***: Visualizations of several clusters based on their geolocation data (left) and activity patterns (right). **a** Geolocations of Cluster 15, **b** Patterns of cluster 15, **c** Geolocations of cluster 17, **d** Patterns of cluster 17, **e** Geolocations of cluster 7, **f** Patterns of cluster 7

income professionals and pursue comfortable, healthy lifestyles with high standards of living.

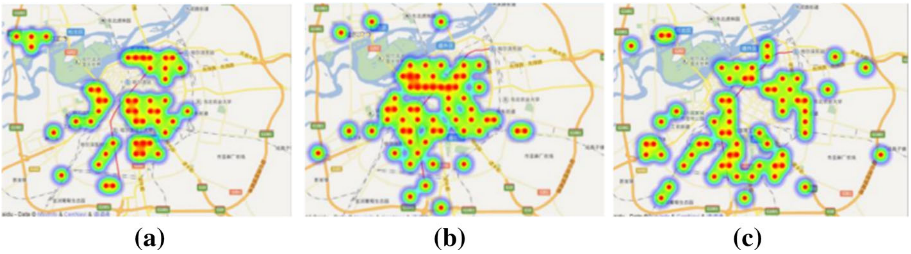


Fig. 8 ***Suning***: Visualizations of the three clusters based on their geolocation data. Their predominant topics are as follows: Cluster 1: Topic 2 (JOMOO and electric stove) and Topic 9 (washing machine, Whirlpool and Little Swan). JOMOO, Whirlpool, and Little Swan are household appliance brands. Cluster 2: Topic 1 (daily necessity, hair care and shampoo), Topic 3 (rice, snack, chocolate and milk), and Topic 7 (toilet paper and home textile). Cluster 3: Topic 4 (Vivo, NOKIA, Smartisan, and HTC) and Topic 6 (air cleaner, SLR camera, Nikon, and Canon). **a** Cluster 1, **b** Cluster 2, **c** Cluster 3

4.4 Summary

LDA effectively discovers topics that represent particular online behaviors. The visualizations of aggregated regions with similar online browsing patterns reveal many meaningful phenomena. First, users in adjacent regions tend to share similar preferences on an online service platform; the browsing patterns of each region are inferred to be related to regional features. Thus, user-oriented recommendations can be made based on the preferences of nearby users or on the regional features of their first POI when considering the cold-start scenario. Second, distinct correlations exist between physical locations and online browsing behaviors with respect to various online services. Users with similar online browsing interests are found in regions with a broader geographical spread for services that are less focused on user locations, whereas they are more geographically confined for an LBS. These findings suggest a strong dependence of online content browsing behaviors on users' first POIs and provide new insights into targeted marketing based on locations. Given the existence of correlations between adjacent regions, the following additional reasonable questions arise: (1) *How are users who share similar online interests correlated with other users located in the same region?* (2) *What are the spatial distributions of users who share similar online browsing interests in a metropolitan city?* To address these questions, we next conduct analyses at the more fine-grained level of individual users.

5 Geospatial Analysis at the User Level

This section describes the analysis of the geospatial properties of users' online browsing behaviors at the level of individual users. We first define *co-poi* and *co-interest* relationships between users and construct two types of networks: *offline co-poi networks* and *online co-interest networks*. To connect users' online browsing behaviors with their physical locations, we then build multilayer networks that support multiple relationships between users. Subsequently, various multilayer-network-based indicators are defined and analyzed. Finally, through comparative analysis, we first answer the question *how are users who share similar online interests correlated with other users located in the same region?* and then address the question *what are the spatial distributions of users who share* *similar online browsing interests in a metropolitan city?* All the discoveries indicate the necessity of location awareness when designing recommendation systems.

5.1 Network Construction

The interpersonal relationships used to create social networks were divided into two types by Kazienko [20]: direct and indirect relationships. The direct relations between users indicate explicit social relationships and reflect mutual acquaintance between users. In contrast, the indirect relations are typically relevant to common interests among two or more users such as two users who comment on the same picture. Here, to describe the relationships between users of online content services, we first define indirect relationships between users such as two users who have the same first POIs or who have browsed items with the same tags. Then, we form an *offline co-poi network* and an *online co-interest network* for each investigated service.

5.1.1 Offline co-poi Network

Based on the *co-poi* relationships, we construct an undirected graph $G^{poi} = (V^{poi}, E^{poi})$ to represent the global similarity among users in terms of their first POIs. V^{poi} is the set of all users $(v_1, v_2, \ldots, v_l)$ in the network, and E^{poi} is the set of edges. If $v_i \in V^{poi}$ and $v_j \in V^{poi}$ possess a *co-poi* relationship, then there is an edge e_{ij}^{poi} connecting them. We call this graph the "*offline co-poi network*". Meanwhile, we calculate the geographic distances D_{ij} between the first POIs fp_i and fp_j of two users v_i and v_j. $D_{ij} = 0$ indicates that two users have a *co-poi* relationship.

5.1.2 Online Co-interest Network

If a user v_i has browsed items with the same tags as those of items browsed by user v_j, we say that these users have a *co-interest* relationship. We use the *co-interest* relationships between users to construct an undirected graph $G^{on} = (V^{on}, E^{on})$, in which an edge $e_{ij}^{on} = (v_i, v_j)$ indicates that two users v_i and v_j have a *co-interest*. Eventually, all users and the links between them will form a network G^{on} that directly reflects the global user interest relations, called the *online co-interest network*.

Table 4 reports the characteristics of the *online co-interest networks* for the three services. ***Youku*** and ***Suning*** have higher average clustering coefficients $\langle CC \rangle$ of 0.82 and 0.90, respectively, whereas the value for ***Meituan*** is 0.56. This result confirms the existence of a small-world effect in these networks [38]: their clustering coefficients are higher

Table 4 Properties of the G^{on} networks for the three services: the numbers of nodes and edges, N and K respectively; the average node degree $\langle k \rangle$; the average clustering coefficient $\langle CC \rangle$; and the average physical distance between nodes $\langle D_{ij} \rangle$ [km]

G^{on}	N	K	$\langle k \rangle$	$\langle CC \rangle$	$\langle D_{ij} \rangle$
Meituan	33,647	4,069,528	1728	**0.56**	8.3
Youku	24,729	3,183,559	763	0.82	12.6
Suning	4086	2,001,355	981	0.90	12.3

than the expected value in a random network of the same size. Because the vertices in a small-world network tend to be clustered into communities, we infer that groups of users with highly similar preferences concerning their item browsing choices exist in these networks. Regarding these users, we will determine the features of their spatial distributions later in this article.

5.2 Multilayer Network Construction

In these networks, parallel interactions exist between users: *co-interest* and *co-poi* relations. Each interaction represents a relation of a different type. An interaction exists between users if and only if at least one relation of either type exists. To consider these parallel interactions, we model the relationships between users using a multilayer network *M*. Unlike normal graphs, which focus on a single-layer abstraction of interactions, a multilayer network explicitly incorporates multiple channels of connectivity to describe systems that contain various types of interconnections [4]. Recently, such networks have become a popular means of modeling complex interacting systems such as biological networks, air pollution, and transportation [4].

Here, we denote a multilayer network consisting of n layers by $M = \{G^1, \ldots, G^n\}$. The α-th ($\forall \alpha \in (1, n)$) layer of the multilayer network is represented by $G^\alpha(V^\alpha, E^\alpha)$, where V^α and E^α are the sets of vertices and edges, respectively, of the graph G^α. A multilayer network that integrates two networks G^α and G^β is illustrated in Fig. 9, along with its three associated link types. A multiplex link is defined as a link that exists in both layers, and the set of multiplex links is denoted by $E^{\alpha\cap\beta}$. A single-layer link is a link that appears in only one layer. The set of all single-layer links on layer α is denoted by $E^{\alpha\setminus\beta}$. In addition, we can associate with layer α an adjacency matrix $A^\alpha{=}\{a_{ij}^\alpha\}$, where $a_{ij}^\alpha = 1$ if nodes i and j are connected through a link on layer α. Such a multilayer network is specified by the vector of the adjacency matrices of the n layers, $A = \{A^1, \ldots, A^n\}$. Suppose that there are N nodes in the multilayer network; then, we define the degree of a node i ($\forall i \in (1, N)$) on a given layer α as $k_i^\alpha = \sum_j a_{ij}^\alpha$, and the degrees of the nodes in the α-th layer of the network form a vector $\vec{k^\alpha} = (k_1^\alpha, \ldots, k_N^\alpha)$.

In accordance with the above definition of a multilayer network, the *online co-interest network* G^{on} and the *offline co-poi network* G^{poi} for each service can be combined to obtain

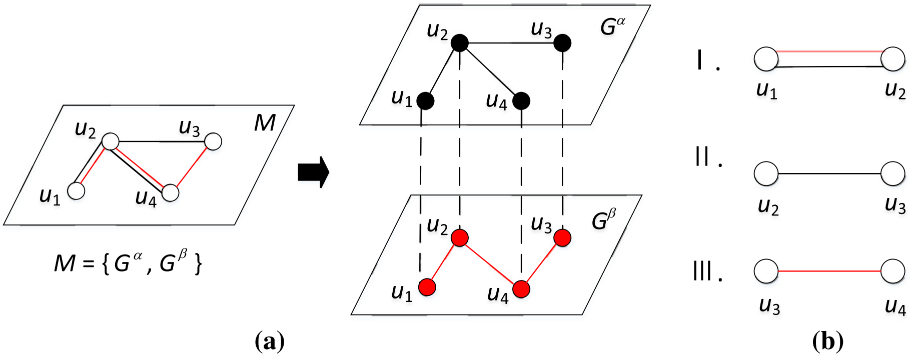


Fig. 9 Multilayer model of a network with I. multiplex links, II. single-layer links on G^α, and III. single-layer links on G^β. **a** Multilayer network, **b** Link types

a multilayer network $M = \{G^{on}, G^{poi}\}$. Thus, in this multilayer network, an edge e_{ij} from user v_i to user v_j exists if at least one relation of either type exists from v_i to v_j. $E^{on \cap poi}$ denotes the set of edges existing in both G^{on} and G^{poi}. Users connected by these edges have both *co-poi* and *co-interest* relationships.

Below, we will define several measures of correlation and spatial distribution based on the multilayer network structure. By applying the same methodology to each of the three services, comparative analyses will be conducted to obtain the general and specific geospatial properties of users' various preferences.

5.3 Correlations Between Co-interest and Co-poi Users

The multi-dimensional social interactions in multilayer networks are commonly described using three widely accepted indicators: the *link overlap*, the *degree correlation*, and the *correlation coefficient per node* [16, 20, 41]. These measures provide complementary insights into the organization of these multi-dimensional interactions. In this section, we apply these measures to mine the correlations between networks composed of *co-interest* and *co-poi* users. The subsequent analyses will answer the question *how are users who share similar online interests correlated with other users located in the same region?*

5.3.1 Measures for Correlation Analysis

Link overlap The *link overlap* metric measures the tendency for links to be simultaneously present in both layers of the multilayer network, namely, the extent to which *co-poi* users also possess *co-interest* relationships. Here, we define the *link overlap* between G^{on} and G^{poi} as follows:

$$overlap = \frac{|E^{on \cap poi}|}{|E^{poi}|} \tag{2}$$

where $|E^{poi}|$ is the number of edges in G^{poi} and $|E^{on \cap poi}|$ is the number of multiplex links.

Degree correlation This measure evaluates the correlation between the degrees of nodes in the two graphs G^{on} and G^{poi}. The *degree correlation* is expressed as follows:

$$\rho(\overrightarrow{k^{on}}, \overrightarrow{k^{poi}}) = \frac{|\overrightarrow{k^{on}} \cap \overrightarrow{k^{poi}}|}{|\overrightarrow{k^{on}} \cup \overrightarrow{k^{poi}}|} \tag{3}$$

where $\overrightarrow{k^{on}}$ is the vector of node degrees in G^{on}. If $\rho(\overrightarrow{k^{on}}, \overrightarrow{k^{poi}})$ is approximately equal to 1, then nodes tend to have equal numbers of neighbors in both G^{on} and G^{poi}, namely, nodes that share similar online browsing interests with many (few) nodes in the network G^{on} also have many (few) *co-poi* relationships in the network G^{poi}.

Correlation coefficient per node The *correlation coefficient per node* represents the extent to which the neighbors of a given node in the two graphs G^{on} and G^{poi} are correlated. We can define the *correlation coefficient* of node i as

$$C_i^M = \frac{\sum_{j=1}^{N} a_{ij}^{on} a_{ij}^{poi}}{\sqrt{\sum_{j=1}^{N} a_{ij}^{on} \sum_{j=1}^{N} a_{ij}^{poi}}} \tag{4}$$

where N is the number of nodes in M and $a_{ij}^{on} = 1$ if nodes v_i and v_j are connected via a link in G^{on}. If C_i^M is approximately equal to 1, then the neighbors of node i in the different layers are nearly identical, namely, the online browsing behaviors of *co-poi* users are alike.

5.3.2 Results of Correlation Analysis

Table 5 presents the values of the three measures, i.e., the *link overlap*, the *degree correlation* and the average *correlation coefficient* for all nodes. Figure 10 depicts the probability distributions of the *correlation coefficients* of all nodes for the three services. According to these results, different services show different levels of correlation between the *co-poi* and *co-interest* networks:

(a) ***Meituan*** This service presents high values for all three measures. The relatively high *link overlap* implies that *co-poi* users are more likely to visit the online records of shops in the same locations. The equally pronounced *degree correlation* indicates that residents with *co-poi* relationships with many (few) neighbors tend to share similar shop-visiting preferences with many (few) individuals. Moreover, ***Meituan*** has a high average *correlation coefficient*. This high average *correlation coefficient* ($\langle C^M \rangle = 0.31$) signifies that 31% of the neighbors of nodes in G^{on} are identical to the corresponding *co-poi* users in G^{poi}.

(b) ***Youku*** & ***Suning*** According to the three metrics, these two services show a similar level of correlation between *co-poi* and *co-interest* users, which is lower than the correlation for ***Meituan***. Since these services have a less prominent location-based focus, this is to be expected. However, their *degree correlations* are still high. The high *degree correlation* values indicate that users tend to have equal numbers of *co-poi* and *co-interest* neighbors. In addition, the relatively high *average correlation coefficient* ($\langle C^M \rangle = 0.2$) indicates that over 20% of *co-poi* users in G^{poi} also share *co-interest* relationships.

5.3.3 Discussion

Considering the correlations between the networks G^{on} and G^{poi}, we analyze the correlations between *co-interest* and *co-poi* users. The three metrics adopted here have also been widely used in other studies. The *link overlaps* of multiple relationship networks for an online game studied by Lambiotte [41] were below 0.2, and the *degree correlations* were smaller than 0.2. Similarly, the *link overlaps* among different types of relations in a film rating dataset [16] have been found to be less than 0.06. Thus, the correlations that we find between G^{on} and G^{poi} are high by comparison, especially for an LBS such as ***Meituan***.

Table 5 The *link overlap*, the *degree correlation*, and the average *correlation coefficient* for all nodes ($\langle C^M \rangle$)

Online content service	*overlap*	$\rho(\overrightarrow{k^{on}}, \overrightarrow{k^{poi}})$	$\langle C^M \rangle$
Meituan	**0.2**	**0.74**	**0.31**
Youku	0.1	0.67	0.20
Suning	0.13	0.67	0.22

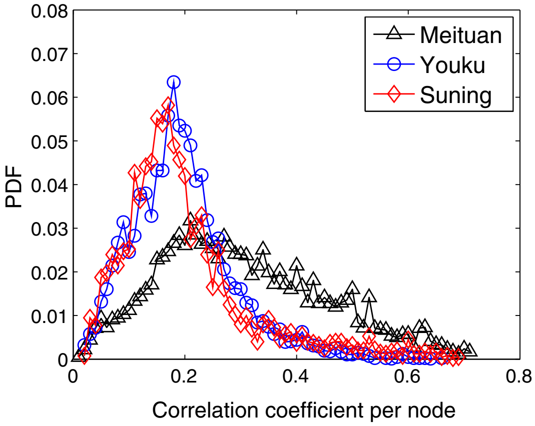


Fig. 10 Probability distributions of the *correlation coefficients* for all nodes

Users tend to have equal numbers of neighbors and even to share the same neighbors in G^{on} and G^{poi}. These phenomena suggest the possibility of mutual link prediction between the networks [18] and the ability to build an effective recommendation system by considering both *co-poi* and *co-interest* relationships. Next, in addition to the *co-poi* relationships, we will infer the overall spatial distributions of users with *co-interest* relationships in a metropolitan city.

5.4 Spatial Distribution of Co-interest Users in a Metropolitan City

Scellato et al. [38] defined two geo-social measures for characterizing how geographic distance affects social structure: a *node locality* metric and a *geographic clustering coefficient*. Here, we extend these two metrics by considering multiple basic network components and relating them to the physical distances D_{ij} between users. Figure 11 shows examples of the three basic network components considered, namely, edges, neighbors, and triads, which reflect distinct dimensions of the relationships among nodes. Our defined metrics will allow us to comprehensively address the question *what are the spatial distributions of users who share similar online browsing interests in a metropolitan city?*

5.4.1 Measures for Spatial Distribution Analysis

Link locality This metric is a measure of the geographic proximity of two users connected by an edge (Fig. 11a) in G^{on}. The *link locality* of edge e_{ij} in G^{on} is quantified as follows:

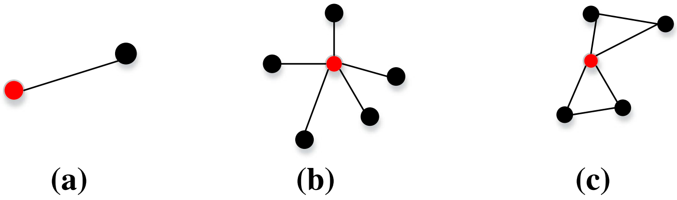


Fig. 11 Typical components of a network. The red dot denotes the target node. **a** An edge of the target node. **b** Neighbors of the target node. **c** Triads containing the target node

$$LL_{ij}{}^{M} = e^{-D_{ij}/\beta} \tag{5}$$

where β is a scaling factor and is defined as the mean distance $\langle D_{ij} \rangle$ between the *fp* of all users in a network. In this way, comparisons can be drawn between different networks containing users with different values of D_{ij}. We also adopt an exponential decay function to emphasize edges spanning shorter geographic distances. By definition, $LL_{ij}{}^{M}$ is always normalized to values between 0 and 1.

Node locality The *node locality* metric quantifies the geographic proximity among the neighbors of a given node and the node itself [38]. Γ_i^{on} denotes the set of neighbors of node i in G^{on}, as shown in Fig. 11b. The node degree k_i is the number of these neighbors, i.e., $k_i = |\Gamma_i^{on}|$. Then, the *node locality* of node i is defined as follows:

$$NL_i^M = \frac{1}{k_i} \sum_{j \in \Gamma_i^{on}} e^{-D_{ij}/\beta} \tag{6}$$

Triad locality Triads, as depicted in Fig. 11c, are sets of three interconnected nodes, which can be interpreted as "the friend of my friend is my friend". The *triad locality* metric quantifies the geographic distances among nodes in triads. A triad composed of nodes i, j and k in G^{on} is denoted by Δ_{ijk}^{on}. Δ_i^{on} denotes the set of triads that contain node i, and the number of triads in Δ_i^{on} is $t_i = |\Delta_i^{on}|$. Then, the *triad locality* of node i in G^{on} can be expressed as follows:

$$TL_i^M = \frac{1}{t_i} \sum_{j,k \in \Delta_i^{on}} \frac{1}{3} \left(e^{-D_{ij}/\beta} + e^{-D_{ik}/\beta} + e^{-D_{jk}/\beta} \right) \tag{7}$$

Geographic clustering coefficient The three metrics above quantify the distances among the first POIs of nodes based on three basic network components. However, it is still essential to investigate the tendency for nodes to cluster together, especially for nodes whose first POIs are in close geographical proximity. The *geographic clustering coefficient* metric is an extension of the clustering coefficient (*CC*) [38]. The *geographic clustering coefficient* of node i is thus defined in the same way as *CC*:

$$GC_i^M = \frac{1}{k_i(k_i - 1)} \sum_{j,k \in \Gamma_i^{on}} w_{ijk} \tag{8}$$

where w_{ijk} is the weight assigned to the triplet of nodes i, j and k and is defined as

$$w_{ijk} = e^{-\frac{\max(D_{ij}, D_{ik}, D_{jk})}{\beta}}$$

Because this measure uses the maximum distance among all links in a triplet, it emphasizes nodes that are all closely related to each other based on their first POIs: when one of the three nodes is not close to the other two, the weight will immediately decrease. A node with a high GC_i^M has tightly interconnected neighbors, and it and both of its neighbors are located in close proximity to each other.

5.4.2 Results of Spatial Distribution Analysis

The city considered in this study has an area of nearly 700 (26 × 26) square kilometers (km), and the average distances $\langle D_{ij} \rangle$ between the *fp* of the users of ***Youku***, ***Suning***, and ***Meituan*** are 12.6, 12.3, and 8.3 km, respectively. Users who employ the local services offered by ***Meituan*** are not as widely distributed as the users of ***Suning*** and ***Youku***, which are services that allow individuals to buy merchandise and watch videos. In addition, the average distances between the *fp* of the *co-interest* users of these three services are 8.8, 8.7, and 6.5 km, respectively, which are much smaller than the $\langle D_{ij} \rangle$ values. These findings indicate that on average *co-interest* nodes are geographically closer to each other than randomly selected nodes. We will later describe in detail how the geographic distances D_{ij} between the first POIs of users affect the distribution of connections in G^{on}.

Link locality The probability distributions of the *link locality* metric are shown in Fig. 12a. The curves peak at a high *link locality* value of 0.7, which corresponds to a physical distance D_{ij} of 3 km for ***Meituan*** or 5 km for ***Suning*** and ***Youku***. By contrast, a fairly small proportion of the edges in G^{on} have a *link locality* of 1.0, namely, a true *co-poi* relationship. Therefore, the *co-interest* users in G^{on} include not only *co-poi* users but also a number of users whose first POIs are separated from each other by a small distance. Moreover, the average values for the three services are all greater than 0.5. Hence, a majority of the links in G^{on} have large *link locality* values, and the first POIs of the users connected by edges tend to be geographically confined.

Node locality The probability distributions of the *node locality* metric are shown in Fig. 12b. The probability initially increases as the *node locality* increases and then exhibits a sharp decline after reaching its peak. As in the case of the *link locality*, the curve for ***Meituan*** is different from those of the other two services. The curve reaches its peak at 0.65. More than 40% of users have *node locality* values of greater than 0.6, and the average value is nearly 0.6. Hence, ***Meituan*** has many *co-interest* users whose first POIs are located within a small geographic distance of each other. A similar effect is observed for ***Youku*** and ***Suning***, but it is weaker than that for ***Meituan***: only 30% of users have a *node locality* of greater than 0.5, and the mean values are slightly above 0.4. The distinct *node locality* features of ***Meituan*** compared with the other two services can be interpreted as follows: People tend to travel limited distances for shopping in the offline world, and thus, the distances between the first POIs of users with similar preferences on group buying services are short. By contrast, the consumption of items offered by services with less of an emphasis on users' locations is influenced by personal interest as well as physical location.

Notably, online social networks exhibit relatively high *node locality* values compared with *online co-interest networks*. Scellato et al. [38] discovered that social LBSs have an average value of over 0.8, whereas social networks that are less focused on user locations

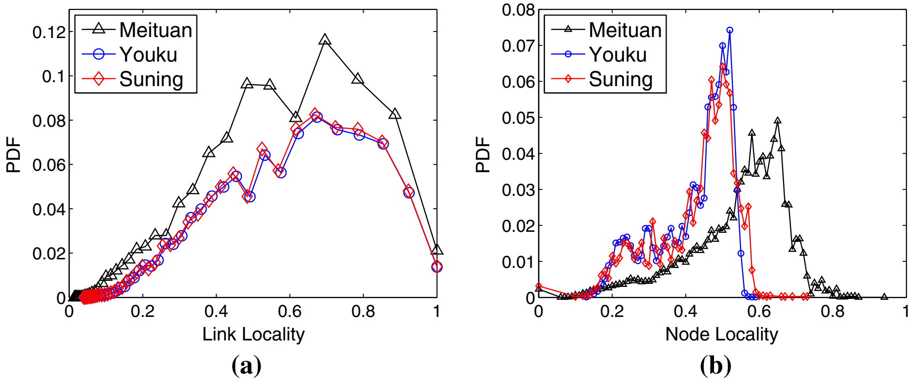


Fig. 12 Probability distributions of **a** *link locality* and **b** *node locality*

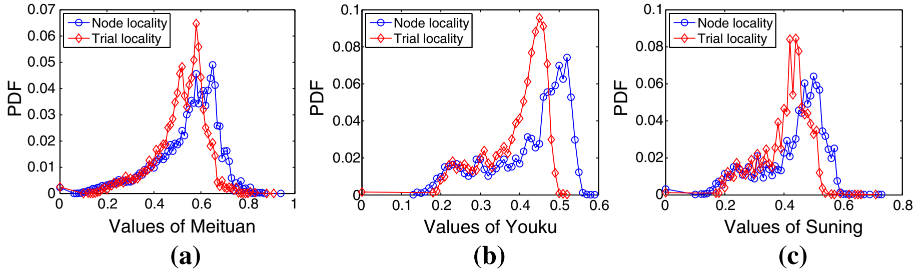
have an average value of 0.5. Human social behavior possesses a tendency toward association with like-minded individuals, leading to the formation of homogeneous social circles both online and offline [17]. By contrast, users' online content browsing behaviors are more strongly affected by personal interests rather than nearby offline individuals. These characteristics cause *online co-interest networks* to exhibit lower values of *node locality*.

Triad locality Figure 13 compares the probability distributions of the *triad locality* and *node locality* for the three services. ***Meituan*** has the largest *triad locality* and shows the most obvious physical proximity among nodes in triads. Overall, the average values of the *triad locality* and *node locality* are similar. The *triad locality* has a narrower distribution compared with the *node locality*: there is no discernible proportion of nodes with markedly lower or higher values of the *triad locality*. This fact reveals that users who are connected in G^{on} with a relatively small or large D_{ij} are not inclined to form a triad with another user. Users who are tightly interconnected because of their similar online interests appear to have more similar physical proximity properties in terms of their first POIs compared with individuals who do not form triads.

Geographic clustering coefficient Here, we study the *geographic clustering coefficients* to understand whether triplets of mutually connected users are more likely to be geographically close to or distant from each other. The three datasets exhibit different values: ***Meituan*** has an average value of 0.394, whereas the mean values for ***Youku*** and ***Suning*** are 0.14 and 0.158, respectively. Note that the standard CC is not affected by the physical distances D_{ij}. Indeed, ***Meituan*** has the lowest $\langle CC \rangle$ but the largest average *geographic clustering coefficient* among the three datasets. These results indicate that the online interest network of an LBS tends to contain more geographically confined triplets than do the networks of services that are focused on commodity buying or personal entertainment. These conclusions provide interpretations of the different shapes of the aggregated groups for ***Meituan*** that are visualized in Fig. 6.

5.4.3 Discussion

In this analysis, we first demonstrated how the distances between users' first POIs affect the structures of G^{on} in multiple dimensions and then analyzed the tendency of users to cluster together with others in close geographical proximity. Common characteristics can be discovered among the three studied networks. First, users who share similar online browsing interests are often users whose first POIs are in close proximity. Hence, recommending items to a user based on the preferences of nearby users may be meaningful for solving the cold-start problem. Taking ***Meituan*** as an example, nearby users can be defined



(a) (b) (c)

Fig. 13 Probability distributions of *triad locality* and *node locality* for each of the three services

as users whose first POIs are located within 3 km of that of the target user, considering that the *link locality* of ***Meituan*** reaches its peak when D_{ij} is equal to 3 km. Second, users who are tightly interconnected because of their similar online browsing interests tend show a more similar spatial distribution of their first POIs. Consequently, a location-aware recommendation system that is specifically designed for users in a particular area may perform better than a system designed for the entire user base. In addition, services that place less emphasis on user locations tend to have similar features, whereas the *co-interest* users of an LBS show more obvious physical proximity correlations. Therefore, the location, as a significant aspect of a recommendation system, is especially critical for LBSs.

Overall, users' locations, especially their first POIs, can serve as bridges between their online and offline behaviors and can broadly affect their online browsing behaviors. The results presented above are thus of great significance for location-aware recommendation systems and can be used to guide the improvement of various services.

6 Conclusion

In this paper, we have related the content browsing behaviors of users to their physical locations and discovered the geospatial properties of usage behaviors for multiple online services. By investigating users of three popular online services in a metropolitan city in China, the online browsing behavior patterns of different geographical regions were visualized and discovered to be similar among adjacent regions. Different types of services were found to show distinct patterns of correlation between the online browsing behaviors of users in adjacent regions. In light of these phenomena, we proposed a multilayer-network-based model to analyze the geospatial properties of users' content browsing behaviors at the more fine-grained level of individual users. Users sharing similar online browsing interests were found to include not only *co-poi* users but also more widely spatially distributed users in adjacent regions. Users who are tightly interconnected with each other, forming closed triangles or triads in the network, tend to have more similar physical proximity properties. Some differences also exist between different types of services: the online browsing interest network of an LBS is strongly geographically confined, whereas services that place less emphasis on locations show a lower physical proximity among *co-interest* users. The results presented in this work have several potential applications, including more effective targeted advertising, more efficient content placement and caching, and faster and more relevant information diffusion.

Our future efforts will be aimed at extracting users' mobility patterns to better quantify the similarity between users, based on which we will conduct further analyses of their geospatial properties. In addition, we will further explore location-aware recommendation systems [29–31] to confirm the geospatial properties of users' online browsing behaviors.

Acknowledgements This work is supported in part by the National Natural Science Foundation of China (61671078, 61601042), Funds of Beijing Laboratory of Advanced Information Networks of BUPT, Funds of Beijing Key Laboratory of Network System Architecture and Convergence of BUPT, and 111 Project of China (B08004).

References

1. Abel, F., Henze, N., Herder, E., & Krause, D. (2010). Interweaving public user profiles on the web. In *International conference on user modeling, adaptation, and personalization* (pp. 16–27). Springer.
2. Baumann, P., Kleiminger, W., & Santini, S. (2013). The influence of temporal and spatial features on the performance of next-place prediction algorithms. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing* (pp. 449–458). ACM.
3. Blei, D. M., Jordan, M. I., Griffiths, T. L., & Tenenbaum, J. B. (2007). Hierarchical topic models and the nested chinese restaurant process. In *NIPS* (p. 2003).
4. Boccaletti, S., Bianconi, G., Criado, R., Genio, C. I. D., Gmez-Gardees, J., Romance, M., et al. (2014). The structure and dynamics of multilayer networks. *Physics Reports*, *544*(1), 1–122.
5. Brodersen, A., Scellato, S., & Wattenhofer, M. (2012). Youtube around the world: Geographic popularity of videos. In *Proceedings of the 21st international conference on World Wide Web* (pp. 241–250). ACM.
6. Calderwood, E., & Freathy, P. (2014). Consumer mobility in the scottish isles: The impact of internet adoption upon retail travel patterns. *Transportation Research Part A: Policy and Practice*, *59*, 192–203.
7. Cha, M., Kwak, H., Rodriguez, P., Ahn, Y. Y., & Moon, S. (2007). I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement* (pp. 1–14). ACM.
8. Chen, Y. F., & Lu, H. F. (2015). We-commerce: Exploring factors influencing online group-buying intention in taiwan from a conformity perspective. *Asian Journal of Social Psychology*, *18*(1), 62–75.
9. Cisco, I. (2012). Cisco visual networking index: Forecast and methodology, 2011–2016. *CISCO White paper* (pp. 2011–2016).
10. Collins, J. L., & Wellman, B. (2010). Small town in the internet society: Chapleau is no longer an island. *American Behavioral Scientist*, *53*(9), 1344–1366.
11. Crane, R., & Sornette, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, *105*(41), 15649–15653.
12. Cranshaw, J., Toch, E., Hong, J., Kittur, A., & Sadeh, N. (2010). Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing* (pp. 119–128). ACM.
13. Dunbar, R. I. M., Arnaboldi, V., Conti, M., & Passarella, A. (2015). The structure of online social networks mirrors those in the offline world. *Social Networks*, *43*, 39–47.
14. Figueiredo, F., Benevenuto, F., & Almeida, J. M. (2011). The tube over time: Characterizing popularity growth of youtube videos. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 745–754). ACM.
15. Holmes, A., Byrne, A., & Rowley, J. (2013). Mobile shopping behaviour: Insights into attitudes, shopping process involvement and location. *International Journal of Retail & Distribution Management*, *42*(1), 25–39.
16. Horvat, E. A., & Zweig, K. A. (2012). One-mode projection of multiplex bipartite graphs. In *IEEE/ACM international conference on advances in social networks analysis and mining* (pp. 599–606).
17. Hristova, D., Musolesi, M., & Mascolo, C. (2014). Keep your friends close and your facebook friends closer: A multiplex network approach to the analysis of offline and online social ties. Eprint Arxiv.
18. Hristova, D., Noulas, A., Brown, C., Musolesi, M., & Mascolo, C. (2016). A multilayer approach to multiplexity and link prediction in online geo-social networks. *EPJ Data Science*, *5*(1), 24.
19. Jain, D., Agrawal, S., Sengupta, S., & De, P. (2016). Prediction of quality degradation for mobile video streaming apps: A case study using youtube. In *International conference on communication systems and networks* (pp. 1–2).
20. Kazienko, P., Musial, K., & Kajdanowicz, T. (2013). Multidimensional social network in the social recommender system. *IEEE Transactions on Systems Man and Cybernetics-Part A Systems and Humans*, *41*(4), 746–759.
21. Kling, F., & Pozdnoukhov, A. (2012). When a city tells a story:urban topic analysis. In *Proceedings of the 20th international conference on advances in geographic information systems* (pp. 482–485).
22. Li, C., & Liu, J. (2016). Large-scale characterization of comprehensive online video service in mobile network. In *IEEE international conference on communications*.
23. Li, S., Qin, Z., & Song, H. (2016). A temporal-spatial method for group detection, locating and tracking. *IEEE Access*, *4*, 4484–4494.
24. Li, S. S., & Karahanna, E. (2015). Online recommendation systems in a b2c e-commerce context: A review and future directions. *Journal of the Association for Information Systems*, *16*(2), 72.
25. Liu, J., Liu, F., & Ansari, N. (2014). Monitoring and analyzing big traffic data of a large-scale cellular network with hadoop. *IEEE Network*, *28*(4), 32–39.

26. Liu, Y., & Sutanto, J. (2012). Buyers purchasing time and herd behavior on deal-of-the-day group-buying websites. *Electronic Markets*, *22*(2), 83–93.
27. Lo, C., Frankowski, D., & Leskovec, J. (2016). Understanding behaviors that lead to purchasing: A case study of pinterest. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 531–540). ACM.
28. Lv, Q., Qiao, Y., Ansari, N., Liu, J., & Yang, J. (2017). Big data driven hidden markov model based individual mobility prediction at points of interest. *IEEE Transactions on Vehicular Technology*, *66*(6), 5204–5216.
29. Ma, Z., Xie, J., Li, H., Sun, Q., Si, Z., Zhang, J., et al. (2017). The role of data analysis in the development of intelligent energy networks. *IEEE Network Magazine*, *31*, 88–95.
30. Ma, Z., & Leijon, A. (2011). Bayesian estimation of beta mixture models with variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(11), 2160.
31. Ma, Z., Xue, J. H., Leijon, A., Tan, Z. H., Yang, Z., & Guo, J. (2016). Decorrelation of neutral vector variables: Theory and applications. *IEEE Transactions on Neural Networks and Learning Systems*, PP(99), 1–15.
32. Maulik, U., & Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(12), 1650–1654.
33. Miranda, F., Doraiswamy, H., Lage, M., Zhao, K., Goncalves, B., Wilson, L., et al. (2017). Urban pulse: Capturing the rhythm of cities. *IEEE Transactions on Visualization and Computer Graphics*, *23*(1), 791–800.
34. Ochi, P., Rao, S., Takayama, L., & Nass, C. (2010). Predictors of user perceptions of web recommender systems: How the basis for generating experience and search product recommendations affects user responses. *International Journal of Human-Computer Studies*, *68*(8), 472–482.
35. Pantano, E., & Priporas, C. V. (2016). The effect of mobile retailing on consumers' purchasing experiences: A dynamic perspective. *Computers in Human Behavior*, *61*, 548–555.
36. Pozdnoukhov, A., & Kaiser, C. (2011). Space-time dynamics of topics in streaming text. In *ACM Sigspatial international workshop on location-based social networks* (pp. 1–8).
37. van Riel, A. C., & Pura, M. (2005). Linking perceived value and loyalty in location-based mobile services. *Managing Service Quality: An International Journal*, *15*(6), 509–538.
38. Scellato, S., Mascolo, C., Musolesi, M., & Latora, V. (2010). Distance matters: geo-social metrics for online social networks. In *Conference on online social networks* (pp. 8–8).
39. Scellato, S., Noulas, A., Lambiotte, R., & Mascolo, C. (2011). Socio-spatial properties of online location-based social networks. In *International conference on weblogs and social media*, Barcelona, Catalonia, Spain, July (p. 2011).
40. Shafiq, M. Z., Ji, L., Liu, A. X., Pang, J., & Wang, J. (2015). Geospatial and temporal dynamics of application usage in cellular data networks. *IEEE Transactions on Mobile Computing*, *14*(7), 1369–1381.
41. Szell, M., Lambiotte, R., & Thurner, S. (2010). Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(31), 636–641.
42. Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, *46*(sup1), 234–240.
43. Wamser, F., Casas, P., Seufert, M., Moldovan, C., Tran-Gia, P., & Hossfeld, T. (2016). Modeling the youtube stack: From packets to quality of experience. *Computer Networks*, *109*, 211–224.
44. Wang, Y. S., Yeh, C. H., & Liao, Y. W. (2013). What drives purchase intention in the context of online content services? The moderating role of ethical self-efficacy for online piracy. *International Journal of Information Management*, *33*(1), 199–208.
45. Xingfu, Y., Pengyi, Z., & Jun, W. (2015). State-behavior modeling and its application in analyzing product information seeking behavior of e-commerce websites users. *New Technology of Library and Information Service*, *6*, 021.
46. Xu, H., Luo, X. R., Carroll, J. M., & Rosson, M. B. (2011). The personalization privacy paradox: An exploratory study of decision making process for location-aware marketing. *Decision Support Systems*, *51*(1), 42–52.
47. Yang, J., Qiao, Y., Zhang, X., & He, H. (2015). Characterizing user behavior in mobile internet. *IEEE Transactions on Emerging Topics in Computing*, *3*(1), 95–106.
48. Yin, H., Cui, B., Chen, L., Hu, Z., & Zhang, C. (2015). Modeling location-based user rating profiles for personalized recommendation. *Acm Transactions on Knowledge Discovery from Data*, *9*(3), 1–41.
49. Yin, H., Sun, Y., Cui, B., Hu, Z., & Chen, L. (2013). Lcars: A location-content-aware recommender system. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 221–229). ACM.

50. Yin, Z., Cao, L., Han, J., Zhai, C., & Huang, T. (2011). Geographical topic discovery and comparison. In *International conference on world wide web* (pp. 247–256).
51. Yu, X., Pan, A., Tang, L.A., Li, Z., & Han, J. (2011). Geo-friends recommendation in gps-based cyber-physical social network. In *International conference on advances in social networks analysis and mining*, Asonam 2011, Kaohsiung, Taiwan, 25–27 July (pp. 361–368).
52. Yuan, N. J., Zheng, Y., Xie, X., Wang, Y., Zheng, K., & Xiong, H. (2015). Discovering urban functional zones using latent activity trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *27*(3), 712–725.
53. Zhang, Y., Tang, J., Yang, Z., Pei, J., & Yu, P.S. (2015). Cosnet: Connecting heterogeneous social networks with local and global consistency. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1485–1494). ACM.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Qiujian Lv is currently a research assistant in the Institute of Information Engineering, Chinese Academy of Sciences. She received her Ph.D. degree from Beijing University of Posts and Telecommunications (BUPT) in 2017. Her current research interests focus on traffic measurement and classification, mobile Internet traffic analysis, cloud computing, and data mining.

Yuanyuan Qiao is currently a lecturer in the School of Information and Communication Engineering, BUPT. She received her B.E. degree from Xidian University in 2009, and received her Ph.D degree from BUPT in 2014. Her research focuses on traffic measurement and classification, mobile Internet traffic analysis and big data analytics.

Yi Zhang is currently pursuing the degree with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China, where he received the B.E. degree in communication engineering in 2016. He is involved in the research of big data mining, and traffic identification and classification.

Fehmi Ben Abdesslem received his M.E. and PhD from the University of Paris 6 in 2008, before working as a research associate at the University of St Andrews, and at the University of Cambridge. He has then been awarded a Marie-Curie research fellowship from the European Commission (ERCIM) to join SICS, and is now a permanent Senior Research Scientist at the Decisions Networks and Analytics laboratory.

Wenhui Lin is currently a senior researcher in Technology Research Institute, Aisino Corporation. He received his B.E., M.E., and Ph.D. degrees from Beijing University of Posts and Telecommunications, China in 2006, 2009, and 2014 respectively. His research focuses on big data analytics, cloud computing and network security.

Jie Yang received her B.E., M.E., and Ph.D. degrees from Beijing University of Posts and Telecommunications, China in 1993, 1999, and 2007 respectively. She is now a professor and deputy dean of School of Information and Communication Engineering, BUPT. Her current research interests include broadband network traffic monitoring, user behavior analysis, big data analysis in Internet and Telecom, etc. She has published several papers on international magazines and conferences including IEEE JSAC, IEEE Trans. on Wireless Communications and IEEE Trans. on Parallel and Distributed Systems. Also, she is the Vice Program Committee Co-Chairs of IEEE IC-NIDC 2014 and 2012.