Taylor & Francis
Taylor & Francis Group

Cartography and Geographic Information Science

ISSN: 1523-0406 (Print) 1545-0465 (Online) Journal homepage: https://www.tandfonline.com/loi/tcag20

A graph-based approach to detecting tourist movement patterns using social media data

Fei Hu, Zhenlong Li, Chaowei Yang & Yongyao Jiang

To cite this article: Fei Hu, Zhenlong Li, Chaowei Yang & Yongyao Jiang (2019) A graph-based approach to detecting tourist movement patterns using social media data, Cartography and Geographic Information Science, 46:4, 368-382, DOI: 10.1080/15230406.2018.1496036

To link to this article: https://doi.org/10.1080/15230406.2018.1496036

CARTOGRAPHY AND GEOGRAPHIC INFORMATION SCIENCE
2019, VOL. 46, NO. 4, 368–382
https://doi.org/10.1080/15230406.2018.1496036

A graph-based approach to detecting tourist movement patterns using social media data

Fei Hu [a,b], Zhenlong Li [c], Chaowei Yang[a] and Yongyao Jiang[a]

[a]NSF Spatiotemporal Innovation Center and Department of Geography and GeoInformation Sciences, George Mason University, Fairfax, VA, USA; [b]Center for Open-Source Data and AI Technologies, IBM, San Francisco, CA, USA; [c]Department of Geography, University of South Carolina, Columbia, SC, USA

ABSTRACT

Understanding the characteristics of tourist movement is essential for tourist behavior studies since the characteristics underpin how the tourist industry management selects strategies for attraction planning to commercial product development. However, conventional tourism research methods are not either scalable or cost-efficient to discover underlying movement patterns due to the massive datasets. With advances in information and communication technology, social media platforms provide big data sets generated by millions of people from different countries, all of which can be harvested cost efficiently. This paper introduces a graph-based method to detect tourist movement patterns from Twitter data. First, collected tweets with geotags are cleaned to filter those not published by tourists. Second, a DBSCAN-based clustering method is adapted to construct tourist graphs consisting of the tourist attraction vertices and edges. Third, network analytical methods (e.g. betweenness centrality, Markov clustering algorithm) are applied to detect tourist movement patterns, including popular attractions, centric attractions, and popular tour routes. New York City in the United States is selected to demonstrate the utility of the proposed methodology. The detected tourist movement patterns assist business and government activities whose mission is tour product planning, transportation, and development of both shopping and accommodation centers.

ARTICLE HISTORY
Received 20 February 2018
Accepted 29 June 2018

KEYWORDS
Twitter; geospatial data mining; tourist movement; big data; Markov clustering

1. Introduction

Understanding tourist movement patterns is important for tourism behavior and management since the patterns reflect a suite of tourism industry activities including public transportation and planning of shopping centers (Mckercher & Lau, 2008). Substantial efforts have been made to map and model the movement of tourists between their home and destinations or among different destinations. Most of the previous studies acquire people's travel data by surveying an individual's location history (Lau, 2007; Lew & Mckercher, 2002; Mckercher & Lew, 2004) or using automatic location-sensing devices, notably GPS (Mckercher & Lau, 2008; Zheng et al., 2011). These data collection methods have two limitations: (1) data are usually collected with high cost but cover a small number of individuals for a certain demographic group in a short time period (Sui, Elwood, & Goodchild, 2012; Yang, Wu, Liu, & Kang, 2017); and (2) the methods are not scalable and cost-efficient to study the tourist movement patterns from a large number of tourists (Zheng, Zha, & Chua, 2012).

With the advances in information and communication technologies, many social media platforms (e.g. Twitter, Flicker, Facebook, YouTube, FourSquare) allow people to publish and share information, images, and videos. Particularly, the adoption of location-aware technologies (e.g. GPS) in the communication devices (e.g. mobile phones, watches) allow stories and information to be shared with their geo-locations on social media (Hu et al., 2015; Yang, Huang et al., 2017). In contrast to travel diary data, social media data are voluntarily generated and usually include a large number of users for extended periods of time. These data not only record the interactions among people and their surrounding environment (Mckercher & Lau, 2008) but capture critical spatiotemporal trajectory features of the social media users (Huang & Wong et al., 2015; Huang & Wong, 2016; Huang et al., 2016).

As a result, social media data with geotagged information provide an alternative data source for many geospatial and social applications (Goodchild, 2007; Sui et al., 2012; Yang, Yu et al., 2017), including analyses of socioeconomic characteristics of social media

CONTACT Zhenlong Li zhenlong@sc.edu

users (Malik et al., 2015), people's references for landmarks and movement patterns (Jankowski et al., 2010), and the flood inundation areas (Huang, Wang, & Li, 2018a, 2018b; Li et al., 2017). Several studies have leveraged social media data to analyze tourism districts, tourist behavior patterns, or provide travel route recommendations (e.g. Kurashima et al., 2010; Lu et al., 2010; Shao, Zhang, & Li, 2017; Zheng et al., 2012). These studies demonstrated the advantages of social media data (e.g. low cost, large volume) on tourism studies. However, these data are big and complex in nature, which makes it hard to quantitatively describe the data and mine the intrinsic knowledge (Sakaki et al., 2010).

To bridge the gap, this paper introduces a graph-based spatiotemporal analysis method to quantitatively detect the tourist movement patterns from social media data. Specifically, a *tourist* model is introduced to identify the social media data generated by the tourists. Second, a DBSCAN-based clustering method is adapted to identify the attractions and edges by clustering the geo-tweets and then constructing the tourist graph. Third, network analytical methods are applied on the tourist graph to detect tourist movement patterns, including popular attractions, most visited point-to-point routes, and centric attractions. Last, a probability graph is constructed to detect the popular tourist routes using the Markov Clustering algorithm (MCL). New York City (NYC) in the United States is used as the study case to demonstrate the proposed approach. The detected tourist movement patterns assist business and governmental entities focused on tour product development, transportation, and development of shopping centers and other accommodations.

2. Literature review

2.1. Activity pattern analysis in social media

Understanding activity patterns contributes to a variety of planning and decision support activities (Noulas et al., 2011). The interaction among population groups has implications for various social and environmental features (e.g. disease spread, hazards, business, culture) (Haythornthwaite et al., 2005; Li et al., 2017; Sakaki, Okazaki, & Matsuo, 2013; Wilson et al., 2009). The key aspect of analyzing activity patterns is to identify the individual spatiotemporal interaction patterns that rely on tracking data for individuals. Travel diary is a common data source to study human activity patterns but is expensive to collect (Li, Goodchild, & Xu, 2013). To simplify the travel diary-type data gathering, GPS devices are widely adopted (e.g. cell phones) but do not cover a large number of users from different social groups. As another data source to support activity pattern studies, social media data are generated by people and allow users to attach geoinformation, allowing the data to be used in activity pattern analysis. For example, Stefanidis, Crooks, and Radzikowski (2013) developed a framework to harvest ambient geospatial information to support situational awareness of human activities. Kisilevich, Mansmann, and Keim (2010) collected geotagged photos to analyze event places by designing a density-based clustering algorithm. Cranshaw et al. (2012) designed a clustering algorithm to map the dynamic urban areas for local activities derived from social media data generated by residents. Gou and Karimi (2017) examined the relationship between human mobility and spatiotemporal features in urban environments. Other studies determined a user's home and work locations from the Twitter data and used the data to examine the individual's activity patterns (e.g. Huang & Wong, 2016; Huang et al., 2016).

Social media data have limitations, notably sparsity and irregularity spatiotemporally (Agichtein et al., 2008). Accordingly, there are several challenges to overcome when analyzing activity patterns using social media data. The first is the demographic bias of user groups in social media platforms being the younger and with positive economic status (Hawelka et al., 2014; Hu et al., 2015; Huang & Wong, 2016; Jiang, Li, & Ye, 2018; Li et al., 2013). Second, social media data with geotagged information take up a small partition of the total published data. For example, about 1% of the public accessible Twitter data has geospatial location information. Moreover, the location information is recorded only when the social media platform is used, so that the data do not reflect the user's activity history in a high percentage (Hasan, Zhan, & Ukkusuri, 2013). Despite these shortcomings, a small percentage of the social media data still can be used to discover interesting spatial mobility patterns (Martin et al., in press; Panteras et al., 2015; Sakaki et al., 2013). Morstatter et al., (2013) state that they were confident that they collected a complete sample of Twitter data when geographic boundary boxes are used for data collection, although the data reflect a small amount of the general pool.

2.2. Tourist movement patterns

According to Haldrup (2004, p. 434), "tourist mobility has often been transformed into a black box explaining the character of specific forms of tourism and tourist behaviour, rather than a phenomenon in its own right that has to be explored and explained." Nevertheless, it is essential to understand movements within a

destination, which can be directly applied to a suite of destination management activities (Mckercher & Lau, 2008).

Twenty-six different itinerary types have been identified by at least five studies, which differ in the mode of transportation, distance, number of stops, and domestic versus international travel (Mckercher & Lau, 2008). The choice of movement patterns depends on the personal power of control and the knowledge of the destination. The spatial distribution of attractions (e.g. clustered, dispersed) influence whether tourists move widely or narrowly within the destination. Mckercher and Lau 2008, p. 363) conducted 1273 arrival interviews to identify 11 tourist movement styles and found them closely related to "territoriality, the number of journeys made per day, the number of stops made per journey, participation in a commercial day tour, participation in extra-destination travel and observed patterns of multi-stop journeys." Mckercher, Shoval, Ng, and Birenboim (2012) utilized GPS data to analyze first and repeat visitor behaviors, which revealed that first-timers and repeaters spend different amounts of time at the same attractions and visit at the different times of the day. Leung et al. (2012) manually collected 500 inline trip diaries to detect the overseas tourist movement patterns in Beijing during the Olympics using the content analysis and social network analytical methods.

The effectiveness of the traditional survey-based methods in the above studies is hindered by issues of cost, scalability, data volume, and privacy. To tackle these issues, researchers turn to social media data. For example, Lu et al. (2010) recovered the existing travel clues from 20 million geotagged photos to suggest customized travel plans according to users' preferences. Zheng et al. (2011) used Flickr data to analyze tourist movement patterns about RoAs, and the topological characteristics of travel routes have been investigated (Zheng et al., 2012). In this work, a sequence clustering method is developed to analyze and distinguish between relaxed versus busy trips. Lee and Tsou (2018) applied the kernel density estimate mapping and dynamic time warping methods on 1-year geotagged Flickr photos in Grand Canyon area to analyze the differences among popular points of interests and spatiotemporal movement patterns of tourists, focusing on hotspot detection and visitation frequency changes by season. Additionally, the interactive visualization of social medial data is useful to understand the movement patterns in social media. For example, Chen et al. (2016) developed an interactive visual analytical system to enable users to select reliable data based on the guidance of a heuristic model and interactive selection tools. However, this method requires human's intervention to detect movement patterns.

Recently with the advance of complex network science, graph-based methods have been applied to quantitatively study tourist movement patterns. Kurashima et al. (2010) built a directed graph model to represent the tourist's traveling sequences and leveraged the topic models with attraction location information to help travelers plan new trips. Schneider, Belik, Couronné, Smoreda, and González (2013) utilized the network theory to detect the daily human mobility patterns, and the experiment results capture ~90% of the population in surveys and mobile phone datasets from different countries. Yang et al. (2017) extended the motif concept in the complex network theory to mine tourist behavior patterns and identify their preferences in an urban environment. Shao et al. (2017) developed a framework to extract and analyze a city's tourism districts from Sina Weibo data.

These previous studies illustrate the capability of graph theories for mining tourist movement patterns at macro- and microlevels, but human intervention is often needed for data collection and cleaning, pattern identification, and extracting ancillary background information of the study activity and area for pattern mining. This paper leverages the power of social media data and graphical analytical theories but focuses more on how to develop a graph-based approach to automatically detect tourist movement patterns at the microlevel from massive social media data with consideration of their preferences. A tourist graph model is proposed and constructed to quantitatively represent a tourist's travel preference among different attractions and detect the tourist's movement patterns by using MCL. The experimental results demonstrate the effectiveness of research approach to the detection of tourist movement patterns and highlight the applicability of the method to other study areas.

3. Methodology

The proposed methods for detecting the tourist movement pattern are composed of three modules (Figure 1). The first collects the geo-tweets in the research area and time and identifies the tweets published by the tourists. The second builds the tourist graph by using DBSCAN algorithm to detect nodes and edges. The third applies the network analysis algorithms (e.g. betweenness centrality, MCLs) to detect tourist movement patterns, including popular attractions, centric attractions, and most popular tour routes. More details of the methodology are elaborated in the following sections.

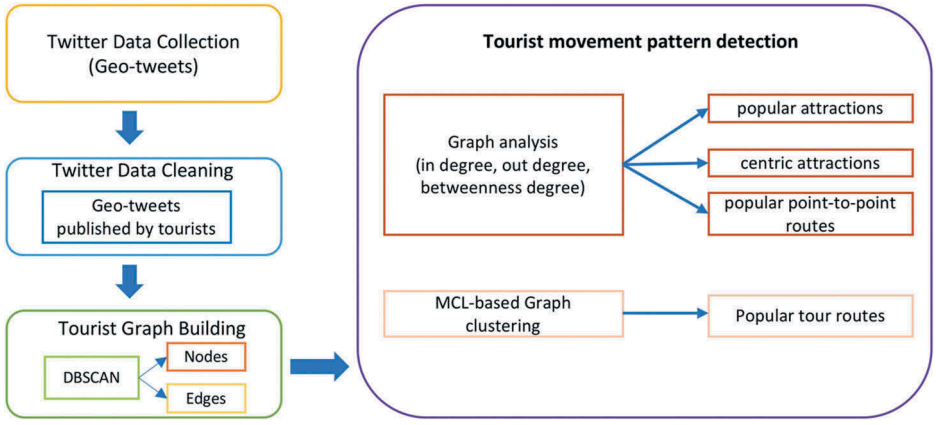

Figure 1. The workflow for the tourist movement pattern detection.

3.1. Tourists extraction

To detect the tourist movement pattern with social media data, the first step is to identify the data generated from tourists rather than from local residents. The existing literature (Li et al., 2013) indicates that a Twitter user can be considered a local resident if the time interval between the first and last tweets in the collected tweets is more than 10 days. Since detecting the movement of a tourist requires at least two location points, the users with only one geotagged tweet during the study time period are excluded. Following the above two rules, the geotagged tweets posted by out-of-town tourists are extracted. The extracted tweets will be further filtered by the DBSCAN algorithm (Section 3.2).

The second step models the tourists and their trajectories by introducing an object-based model for the target social media (Twitter) user called "*tourist*". A *tourist* is defined as follows:

$$tourist = \{id, (x_1, y_1,\ t_1), (x_2, y_2,\ t_2), \ldots, (x_n, y_n,\ t_n)\},$$

where by each *tourist* contains an *id* and a list of tweets. Each tweet, denoted as $(x_1, y_1,\ t_1)$, stores information about time and location (latitude or x, longitude or y); n is the total number of tweets a user posts within the study area and time. Specifically, the metrics are $n \geq 2,\ t_n - t_1 \leq 10\ days,\ t \in TimePeriod,\ (x, y) \in StudyArea$.

Based on the *tourist* model and geotagged tweets, a spatiotemporal trajectory is generated to show the tourist's journey path. As an example, a tourist's trajectory (blue lines, Figure 2) is displayed in a space-time cube with the x-, y-, and z-axes representing latitude, longitude, and time dimension, in series. The green dash line is the 2D trajectory decomposed from the 3D trajectory with only latitude and longitude dimensions.

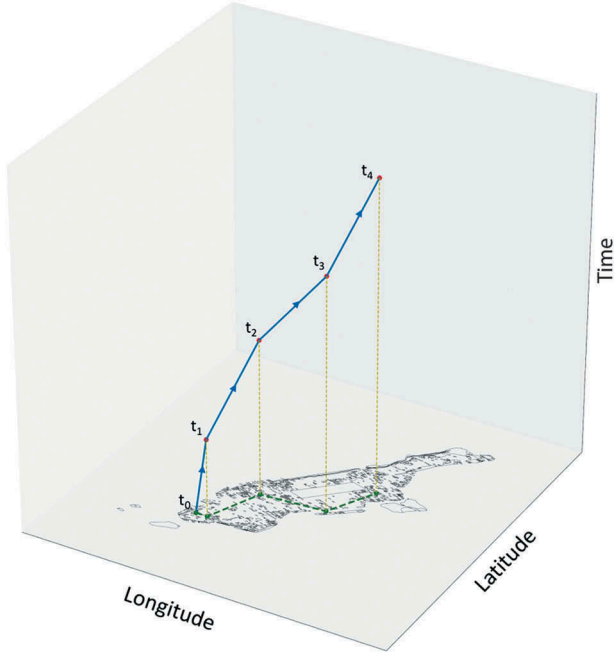

Figure 2. Illustration of the spatiotemporal trajectory for a single tourist (blue line).

3.2. Tourist graph construction with DBSCAN clustering

To analyze the movement patterns from the massive tourist data extracted earlier (Section 3.1), a tourist graph with a set of vertices and edges is built to conduct the spatiotemporal network analytics. On the tourist graph, the vertex represents the attraction, and the edge connecting two vertices indicates there are tourists moving among these two attractions (Figure 3). To define the tourist graph, let V be the vertex of the graph, and E be the edge of the graph, so that the graph G is represented as $G = (V,\ E)$. The vertex set of graph G is denoted as $V(G) = \{v_1,\ v_2, \ldots\ldots,\ v_n\}$, and the edge set is denoted

as $E(G) = \{e_1, e_2, \ldots\ldots, e_m\}$. Each edge $e \in E$ connects two adjacent vertices, *u and v*, which are the edge's start and end points, respectively.

Based on the graph definition, the first step in building the tourist graph identifies the representative vertices (place of attractions). The second step generates the edges between the vertices. The vertices should be based on a common agreement by different tourists and have a significant number of tourists who published their tweets nearby. However, the locations in geotagged tweets are resolved to six decimal places of latitude and longitude, equivalent to 10 cm. This resolution means that even for the same building, the location for different tweets is likely to be different. To address the location variance for the place of attraction, the points visited by the tourists is clustered to detect the graph vertex.

Density-based spatial clustering of applications with noise (DBSCAN) is a density-based clustering algorithm (Ester et al., 1996). The DBSCAN has several advantages in clustering massive social media data (Hu et al., 2015) to extract the vertices of the tourist graph. In the first advantage, the two parameters *Eps* (search radius) and *MinPts* (minimum number of points in the search radius) match well with the definition of vertex in the tourist graph so that it allows the locations to be in a certain range, while the cluster needs to meet the density threshold. In the second advantage, the shape of clustering results can be arbitrary, so it represents the attractions with different shapes. In the third advantage, it does not require the number of clusters in advance (as it is unknown how many vertices or attractions exist before clustering).

The key to getting the appropriate clustering results from DBSCAN is setting the proper *Eps* and *MinPts*. In general, a larger *Eps* generates a broader cluster, while a smaller *Eps* produces a smaller region. The *MinPts* determines the number of clusters: a larger *MinPts* creates a cluster with higher significance but excludes some interesting areas, while a smaller *MinPts* generates more clusters but includes more noise points (Hu et al., 2015; Huang & Wong, 2016). Herein, it is proposed that the two parameters be selected based on the actual study area and the data to be used (Section 5.1).

Once the graph's vertices are identified, the points in the same cluster are updated to be the centroid coordinate of the cluster. Meanwhile, the points outside of the clusters are treated as noisy data to be further filtered (Figure 3). The hollow circle represents the original locations of the tourists' tweets, the bigger solid circle represents the centroid coordinate of the cluster (i.e. vertex in the tourist graph), and the gray circle represents the filtered noise points. The next step constructs the edges between the vertices. As the gray lines illustrate in Figure 3, the tweets published by each tourist are sorted by publishing time, and then the edge

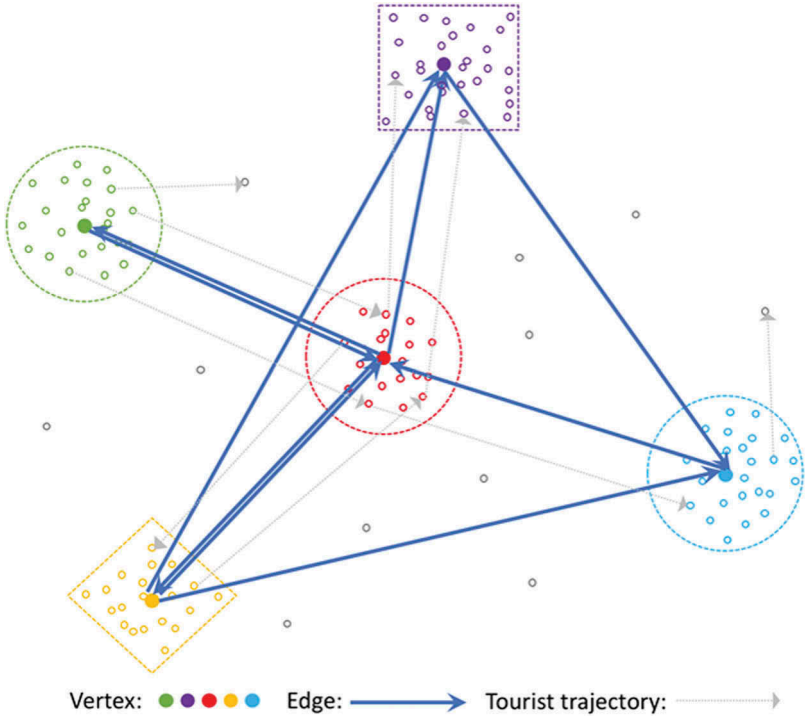

Figure 3. Illustration of the tourist graph construction. (Note that the dash circles and dash squares represent the boundary of the clustering results. The different shapes between them mean that the clustering result may be in different shape with each other).

is connected from the tweets published earlier to the one at the next time stamp until the tourist's last tweet. This operation is repeated for all the tourists, resulting in repeated edges from multiple tourists. The number of repeated edges is set as the weight of the edge. It is noted that the noisy points outside of the clusters are discarded when constructing the tourist graph, which further filters tweets not published by the tourists. The proposed tourist graph provides a graph-based model to quantitatively summarize the spatial transition/movement flows from massive and noisy social media data and builds a bridge for applying the graph network analysis methods to discover the activity patterns for social media users.

3.3. Tourist movement pattern detection

Popular attractions and routes are often those with the largest number of tourist visit. Meanwhile, the tourist traffic flows usually gather at a certain few attractions (i.e. centric attraction). Herein, the tourist movement patterns are quantitatively measured from the perspectives of popular attractions, point-to-point routes, centric attractions, and popular routes.

3.3.1. Detecting the popularity of attractions

The popular attractions (i.e. most visited point-to-point routes, centric attractions) are analyzed based on the constructed tourist graph.

The *popular attractions* are detected by measuring the weighted degree of a vertex in the tourist graph. The weighted degree for each node (s_i)including the in-weighted degree (s_i^+) and out-weighted degree (s_i^-) is calculated using Equations (1), (2), and (3), where $i,\ j, and\ k$ are the vertex id $\in [0,\ n]$, w_{ji} is the weight for the edge from the vertex v_j to the vertex v_i, and w_{ik} is the weight for the edge from the vertex v_i to the vertex v_k. The rank of the nodes by the weighted degree indicates the attraction's popularity.

$$s_i^+ = \sum_{j=1}^{n} w_{ji} \tag{1}$$

$$s_i^- = \sum_{k=1}^{n} w_{ik} \tag{2}$$

$$s_i = s_i^+ + s_i^- = \sum_{j=1}^{n} w_{ji} + \sum_{k=1}^{n} w_{ik} \tag{3}$$

The *most visited point-to-point routes* are detected based on the weight of the edges once the tourist graph is constructed. The weight (w_{ij}) for the edge that connects the vertices v_i and v_j is indicated by how many times this edge is visited.

The *centric attraction* is the place where tourist traffic tends to flow to or leave from and is discovered by measuring the betweenness centrality of the tourist graph. The location of centric attractions may be affected by popularity, geographical location, and transportation convenience (Zheng et al., 2012). The betweenness centrality $C_B(v_m)$ for the vertex v_m indicates the vertex's centrality in a network and equals the number of shortest paths from all vertices to all others that pass through the vertex v_m (Brandes, 2001). The betweenness centrality is computed as follows:

(1) for each pair of vertices v_i and v_j, compute the shortest paths and denote as σ_{ij};
(2) For the above shortest paths, count the number of paths that pass through the vertex v_m and denote as $\sigma_{ij}(v_m)$; and
(3) Compute the fraction of shortest paths that pass through the vertex, and sum the fractions for all pairs of vertices (Equation (4)).

$$C_B(v_m) = \sum_{i \neq m \neq j \in V} \frac{\sigma_{ij}(v_m)}{\sigma_{ij}} \tag{4}$$

Note that the shortest path is the one with the highest sum of edge weights (i.e. path visited by most people). Therefore, a vertex with a high betweenness centrality has a significant influence on the spatial movement patterns of tourists.

3.3.2. Popular tour routes detected with MCL

To detect the popular tour routes, the weighted graph $G_w(V,\ E,\ w(u,v))$ (Section 3.2) is used to compute the transition probability graph of the tourists, where V denotes the vertex of the graph, E denotes the edges of the graph, and $w(u,v)$ denotes the weight matrix which measures the frequencies of the tourists traveling from the vertex u to the vertex v.

Based on the weighted graph, the transition probability matrix $M_{p(u,v)}$ for the graph G_w is derived, where $p(u,\ v)$ denotes the probability of the vertex v being the next stop for a tourist at the vertex u. By introducing the probability matrix, the tourist graph G is further expressed as follows:

$G_p = (V,\ E,\ p(u,v))$, where

$$p(u,v) = \frac{w(u,v)}{\sum_{i=0}^{n} w(u,\ o_i)},\ o_i = \{v \in V \mid w(u,\ v) \rangle 0\} \tag{5}$$

Based on the probability graph G_p, we are able to estimate how likely a tourist travels from one to another connected vertex. If the tourists follow the

transition probability graph to randomly walk among vertices, it is possible to discover at which vertices the tourists tend to gather. The frequently gathered places are considered as popular tourist routes revealed from the collected tweets. Random walks on the proposed graph are performed by adapting "Markov Chains" (Hastings, 1970) using the probability matrix $M_{p(u,v)}$ as the initial input. This assumes that the probabilities for the next time step only depend on the current probabilities. Specifically, three operations are performed in the MCL to detect the popular tour routes from the tourist graph as follows:

(1) The expansion operator (Equation (6)) allows the tourists to connect different regions of the graph by taking the Markov Chain transition matrix powers. A large expansion rate expedites clusters merging. For the expansion rate = 2, the expansion equation is as following

$$Expa(M_{p_{ij}}) = \sum_{k=1:n} M_{p_{ik}} * M_{p_{kj}} \tag{6}$$

(2) The inflation operator (Equation (7)) strengthens the strong parts and weakens the weak parts of the transition probability matrix, and the existing clusters are strengthened by a big inflation rate (r); and

$$Infl(M_{p_{ij}}) = \frac{(M_{p_{ij}})^r}{\sum_{j=1}^{k} (M_{p_{ij}})^r} \tag{7}$$

(3) The repeat expansion and inflation operators lead to the steady state of the probability graph with the clustered vertices and Equation (8) is used to identify if the probability graph reaches the steady state as follows:

$$Diff\left(M_{p1} - M_{p2}\right) = \sum_{i=1:n} \sum_{j=1:n} M_{p1_{ij}} - M_{p2_{ij}} \tag{8}$$

The resultant clusters represent the popular routes that tourists prefer to visit, and the vertices (attractions) have some similar natural or cultural value or offer similar leisure, adventure, and amusement. Herein, the grid hyper-parameter search is conducted with different pairs of expansion ([1, 8]) and inflation rates ([1, 8]) to identify the best pair of expansion and inflation rates. When the expansion and inflation rates become larger, the steady state is achieved faster. However, more vertices are discarded in the final clustering results and each cluster contains fewer vertices because the larger expansion and inflation rates greatly weaken the connection between the preferred and less popular vertices. Based on the grid hyper-parameter search results, both the expansion and the inflation rates are set to 2 in the MCL algorithm to keep the detected vertices in the clustered results, since all these vertices are detected as popular attractions. The selected pair of expansion and inflation rates means that after the tourists visit an attraction, the probability for other attractions is inflated by taking an inflation operator.

4. Case study and data

To demonstrate how the proposed approach detects tourist movement patterns, New York City (NYC) is selected as a case study because it is one of the most tourist-frequented cities in the United States, attracting 62.8 million tourists in 2017 (NYC & Company, 2018).

Geotagged tweets (geo-tweets) of the NYC metropolitan region are collected using the Twitter streaming API from 1 July 2016 to 30 April 2017. The total number of tweets is 5,019,637, and each tweet contains the user id, latitude, longitude, date, text, and other information. These geo-tweets are displayed as a density map (Figure 4) with color brightness representing tweet density (the brighter, the higher density). Moreover, the top-50 attractions in NYC ranked by TripAdvisor (https://www.tripadvisor.com), the world's largest travel site in the world, are manually extracted to evaluate the results (blue circles, Figure 5).

5. Results and discussion

5.1. Vertices detected using DBSCAN and the constructed graph

Based on the *tourist* model (Section 3.1), 55,957 tourists who posted 258,540 geotagged tweets from the NYC Twitter dataset were extracted after which the DBSCAN clustering algorithm (Section 3.2) was applied on the extracted geotagged tweets to filter noisy tweets and identify the tweet clusters (attractions). This reduces the number of geotagged tweets to 109,998. To evaluate the accuracy of the detected tourists, 100 tourists were selected from the detected tourists to manually check their twitter profile and posted tweets. Of these, 94 users are identified as out-of-town tourists based on the location information in their profiles and the locations of their historical tweets and only two are residents in NYC. Four users are company accounts or nonhuman. Accordingly, the detection accuracy of tourists is 94%.

Using DBSCAN clustering, the centroid of each cluster is the vertices of the tourist graph. Herein *Eps* (search radius) is set as the mean length of the street block on the Manhattan Island (~100 m) because these attractions can be well separated at the block scale. To find the appropriate *MinPts*, a grid hyper-parameter

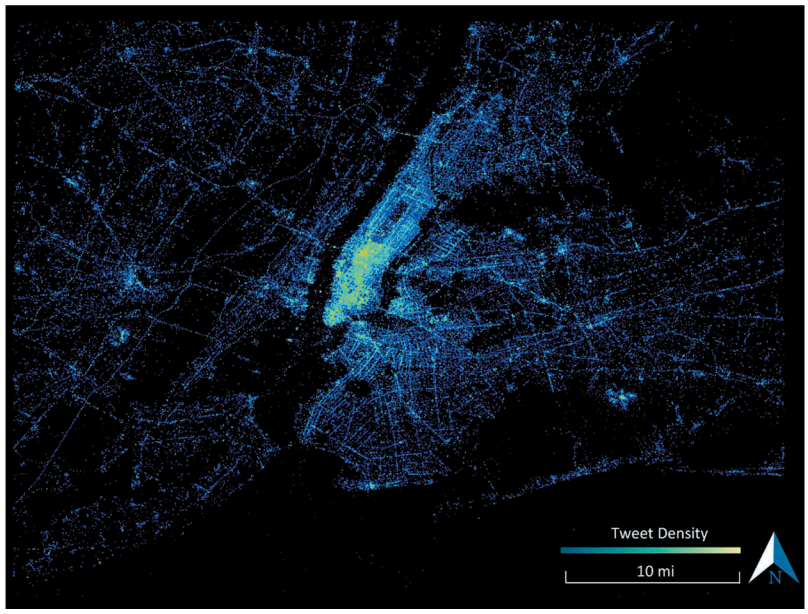

Figure 4. The NYC study area and geotagged Twitter data.

searching approach is selected wherein it is initially set to be from 100 to 5000 with a 100 step and then comparing the different clustering results with the distribution of the top 50 places of interest obtained from the TripAdvisor. Based on the experiment results, the minimum number of points in the search radius (*MinPts) is set* to 1500.

The clustering result with 16 clusters, where the orange icons represent the centroid point for each cluster and the blue circles represent the top 50 attractions from TripAdvisor is illustrated (Figure 5). The centroid of each cluster is selected as the vertex of the graph and labeled by one of their most nearby attractions listed on TripAdvisor. The labels for each vertex and the number of points within the cluster that vertex represents are provided (Table 1).

After connecting the edges between vertices based on tourists' traveling trajectories, the final graph with the nodes and edges is constructed (Figure 6) in which the arrow indicates the direction from one vertex to another and the thickness of the edges represents the frequency of the edge visited by tourists.

5.2. Tourist movement patterns

5.2.1. Popular attractions

The concept of weighted degree indicates the popularity of an attraction (Section 3.3.2). The weighted degree for the 16 nodes (clusters) (Figure 7) and the weighted degree for each node by the font size of labels (Figure 8) is illustrated. The top 10 attractions are Times Square, World Trade Center, Top of the Rock Observation Deck, Brooklyn Bridge, Central Park, Empire State Building, The High Line, MoMA, The Metropolitan Museum of Art, Statue of Liberty, and Madison Square Garden. Their rankings at TripAdvisor are 27, 14, 4, 9, 1, 11, 8, 2, 3, 15, and 23 in series. All of the identified popular attractions are ranked as top 25 by TripAdvisor, and 60% of our results are at the top 10 attractions ranked by TripAdvisor. It is proposed that the difference is due to the different ranking approaches as the research methodology ranks based on how many tourists visit, whereas TripAdvisor considers every reviewer's evaluation score.

5.2.2. Centric attractions

The centric attractions are the places tourists prefer to visit when moving from one attraction to another. The betweenness centrality is calculated to identify the centric attractions. The centric attractions (Figure 9) highlight the betweenness centrality by the label size of the node. The nodes with the largest betweenness centrality are Times Square, World Trade Center, the Museum of Modern Art (MoMA), Metropolitan Museum of Art, Brooklyn Bridge, the High line, and Ground Zero Memorial, all physically close to metro, bus stations, or the traffic centers. These seven nodes are evenly distributed across the Manhattan Island without clustering and are the centric regions of attractions and treated as transition traffic centers. This finding suggests that allocating more transportation resources to these seven regions would help relieve traffic pressure.

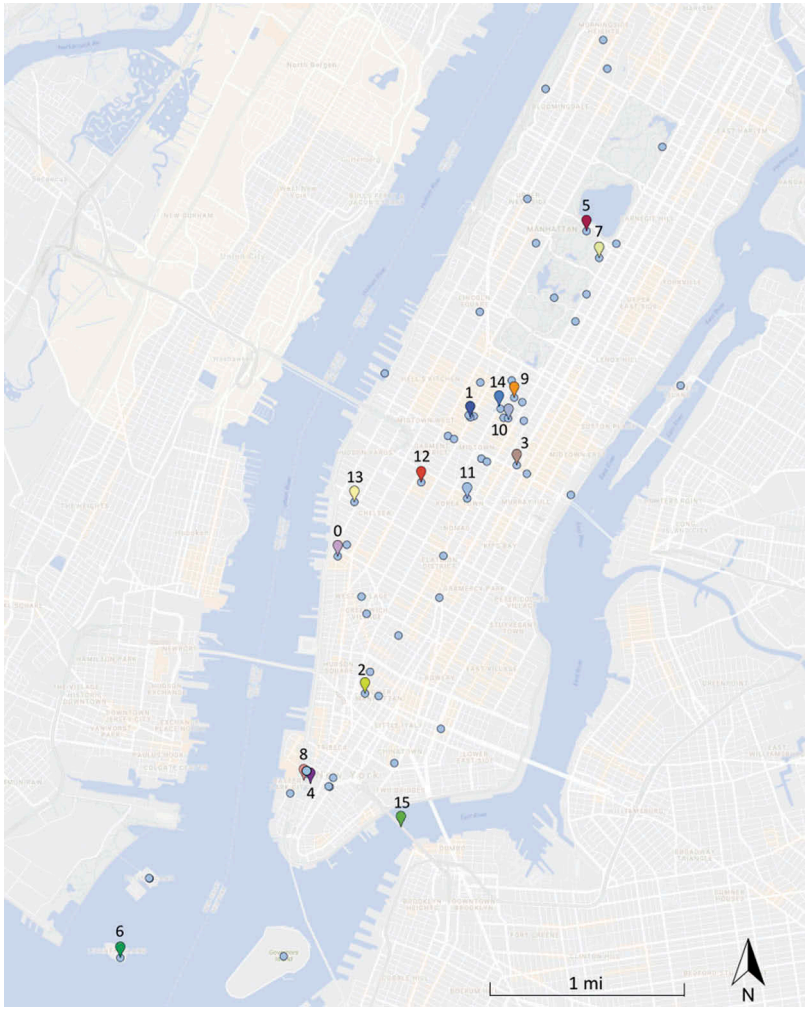

Figure 5. The clustering result using DBSCAN.

5.2.3. Popular point-to-point routes

The popular point-to-point routes are identified by the weighted degree for each node (Figure 10), where the width and color of edges indicate the weighted degree for each node. The top three routes are identified as being (1) from Top of the Rock Observation Deck to Times Square, (2) from Times Square to World Trade Center, and (3) from Central Park to Times Square. All three point-to-point routes contain Times Square as it has the highest weight. Moreover, the distance of the three routes differs among the three, indicating that the distance is not the key factor for trip planning in NYC.

5.3. The popular tour routes

Using the Markov clustering approach, the detected clustering vertices based on the probability graph G_p (Table 2) identify the most probable tourist movement patterns from tourists' preferences.

Two clusters of vertices (attractions) are detected (Table 3) and the attractions in each cluster are connected using the shortest path (Figure 11). The first cluster is centered at the Grand Central Terminal with eight members consisting of museums and historical attractions (Figure 11(a)). The second cluster is centered at the Times Square with eight members themed with modern architecture (Figure 11(b)). When examining the two cluster members in the spatial context, it is evident that members of both clusters are not geographically clustered, indicating that the geographic distance in NYC may not be the determining factor for the tourist travel pattern. Conversely, the tourists' preference for the attractions plays a more critical role in determining the tourist movement patterns. Lastly and based on the clustering results, the two most popular tour routes (Figure 11) are identified and recommended for tourists visiting NYC. For the tourists identified herein, 38.1% visited the attractions in the first cluster, 44.2% visited the attractions in the second cluster, and 17.7% visited both the attractions.

Table 1. The selected attractions used for labeling each vertex.

ID	Coordinates	Label	Number of points
0	40.740272,-74.008383	Meatpacking District	4286
1	40.758895,-73.985131	Times Square	22,843
2	40.723382,-74.003198	SoHo	2914
3	40.752735,-73.977002	Grand Central Terminal	5484
4	40.711246,-74.012749	Ground Zero Memorial	4071
5	40.768924,-73.975307	Central Park	6965
6	40.689189,-74.044654	Statue of Liberty	5968
7	40.77913,-73.962974	The Metropolitan Museum of Art	6000
8	40.711805,-74.012641	World Trade Center	20,672
9	40.761478,-73.977123	The Museum of Modern Art (MoMA)	6157
10	40.758985,-73.979257	Top of The Rock	9853
11	40.748803,-73.985626	Empire State Building	6393
12	40.750465,-73.993521	Madison Square Garden	5029
13	40.744263,-74.006199	The High Line	6181
14	40.759873,-73.978917	Radio City Music Hall	3530
15	40.706086,-73.996864	Brooklyn Bridge	8432

As a further validation, the research results for the discovered routes are compared with the Big Bus Tour Routes in NYC (https://www.bigbustours.com/en/new-york/new-york-routes-and-tour-maps/). The first clustered attractions are similar to that of the downtown and uptown tour of the Big Bus Tours (Figure 11(a)), and the second clustered attractions are similar to that of the midtown and uptown tours (Figure 11(b)). This similarity further supports the utility of the MCL-based probability graph clustering approach in helping to discover the most popular tourist routes.

6. Conclusion and future research

A graph-based approach is introduced to detect the tourist movement patterns from massive and noisy social media (Twitter) data, and an object-based model is designed to represent the tourist's spatiotemporal movement trajectory. To build the tourist graph, the DBSCAN-based method is used to cluster the tourist trajectories to identify the vertices in the graph and then connect the vertices by using the tourist trajectories to generate the edges of the graph. Once the tourist graph is constructed, a set of graph-based network analysis methods is introduced to detect the most common tourist movement patterns.

New York City is used to evaluate the proposed approach. The tourist movement patterns are identified by detecting the popular attractions, centric attraction, popular point-to-point routes, popular tour routes from the tourist graph, and the results demonstrate that the proposed methodologies provide a feasible and effective way to build a graph-based network model for tourists from big social media data. The

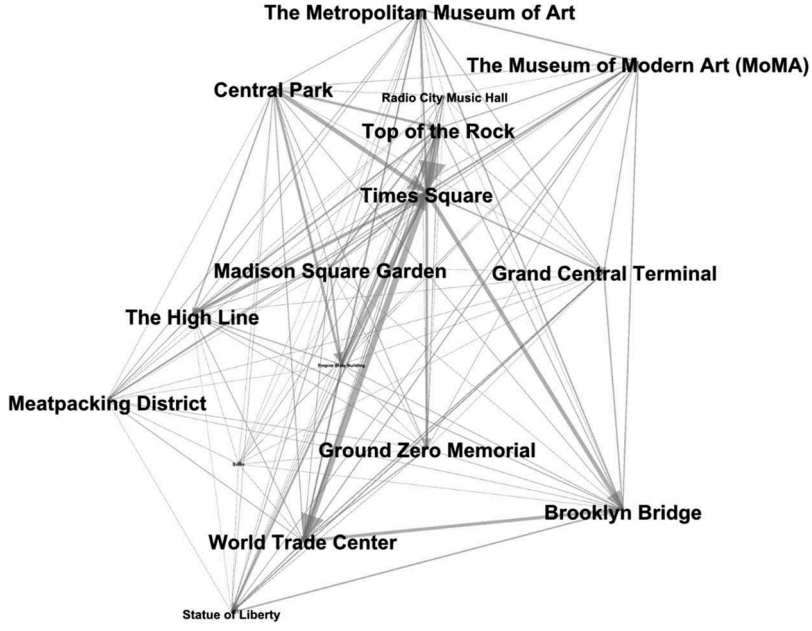

Figure 6. The network graph for the tourists in NYC from 1 July 2016 to 30 April 2017.

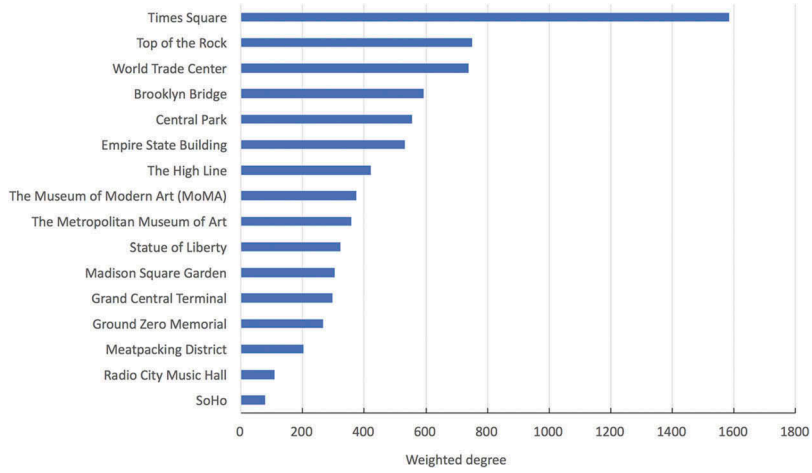

Figure 7. The weighted degree for each node.

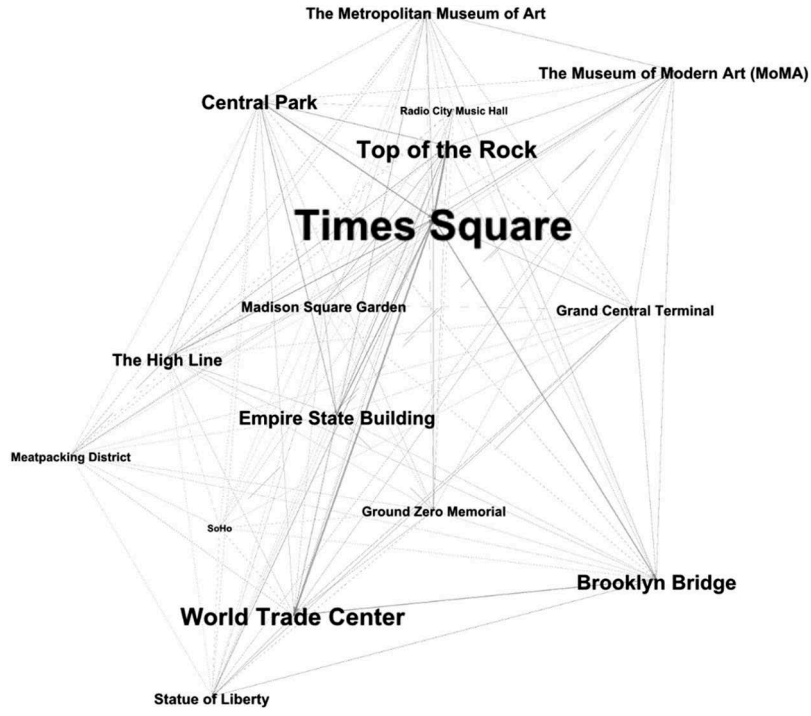

Figure 8. The graph visualization by the weighted degree (font size of labels indicates the node weighted degree).

proposed graph-based methodologies can be extended to study other topics in human dynamics studies in large areas using social media data. Compared with the traditional tourist behavior studies, this study proposes a cost-efficient approach to automatically identify the behavior features of tourists from the big and noisy social media data (geotagged tweets) using quantitative methods.

While the results are promising, several limitations are recognized, and more efforts are needed to improve the approach. First, while the detection accuracy of tourists is high (94%), a small percentage of bots remained in the identified tourists. Fully eliminating these bots in an automatic manner without manual checking is challenging (Guo & Chen, 2014). A potential new line of research is to leverage more advanced methods, such as artificial intelligence, natural language processing, and image recognition to improve the understanding of the social media data content and the accuracy of noisy data filtration and bot detection.

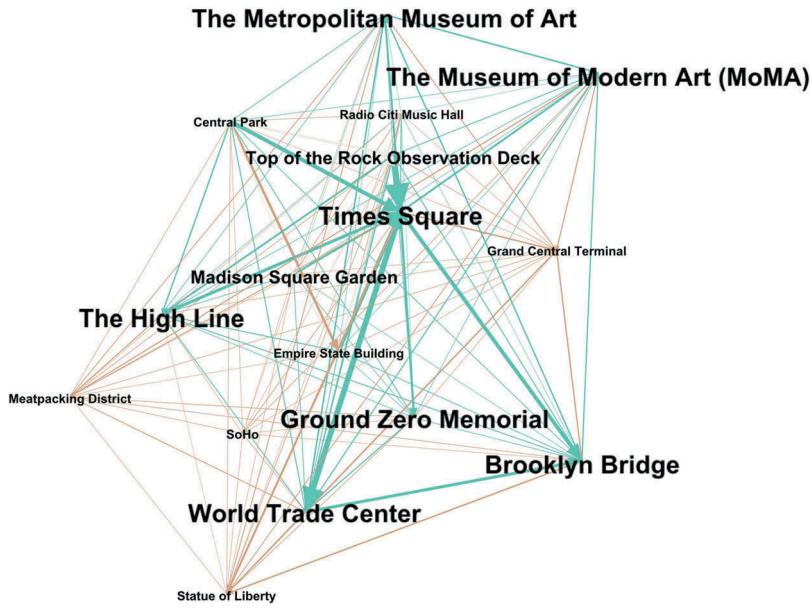

Figure 9. Visualization of the betweenness centrality showing the centric attractions (label font size indicates the betweenness centrality).

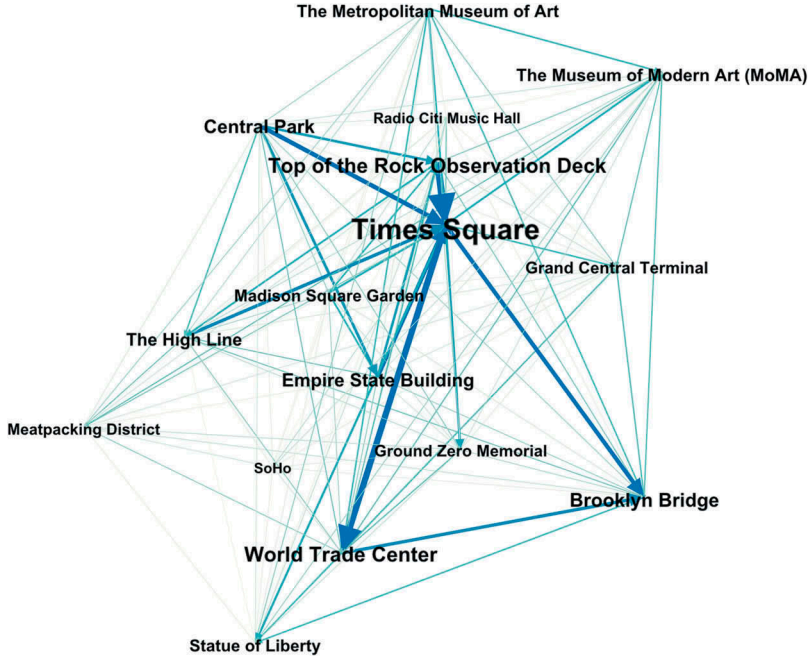

Figure 10. Visualization of the weighted degree among nodes (label font size indicates the weighted degree between nodes).

Second, the trajectory derived from the tweets may not represent the actual trip trajectory as the tourists may not publish tweets for every visited attraction or the tweets were not streamed into our dataset. Third, the movement sequences along the time dimension is another import attribute of the tourist behaviors, which reflects how tourists choose the next stop based on the previous stops. Limiting the graph to only consider the spatial aspect may hide the underlying detailed tourist movement patterns in the time series. Future research should utilize time-series-based methods to further improve the tourist graph by computing the movement probability in considering the individual's previous movements. Lastly, only one data source is considered, and the fusion of multiple data sources (e.g. Flickr, Uber, Facebook) may help discover more interesting features and patterns of tourist behaviors.

Table 2. The probability matrix calculated from the weight matrix for the tourist graph.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0.0000	0.0810	0.1750	0.0670	0.0560	0.0360	0.0160	0.0280	0.1180	0.0470	0.0500	0.0250	0.0160	0.0120	0.1140	0.1590
1	0.0720	0.0000	0.0230	0.0040	0.0920	0.0970	0.0210	0.1040	0.1810	0.0590	0.1380	0.1190	0.0140	0.0230	0.0300	0.0220
2	0.0720	0.0730	0.0000	0.0320	0.0480	0.0520	0.0240	0.0440	0.1770	0.0460	0.0800	0.0540	0.0200	0.0440	0.1060	0.1270
3	0.0510	0.1450	0.0090	0.0000	0.0760	0.1070	0.0000	0.0000	0.0200	0.0800	0.4410	0.0000	0.0000	0.0020	0.0680	0.0000
4	0.0250	0.0450	0.0390	0.0550	0.0000	0.1600	0.0490	0.0820	0.0000	0.0000	0.0040	0.4520	0.0000	0.0000	0.0650	0.0250
5	0.0040	0.0710	0.1090	0.0080	0.0160	0.0000	0.0350	0.0610	0.3020	0.1450	0.0430	0.0590	0.0280	0.0340	0.0420	0.0440
6	0.0140	0.0730	0.0080	0.0450	0.0320	0.0460	0.0000	0.0310	0.2480	0.0280	0.0700	0.1090	0.1010	0.0000	0.0240	0.1710
7	0.0480	0.4540	0.0180	0.0410	0.0220	0.0450	0.0690	0.0000	0.0140	0.0400	0.1290	0.0080	0.0450	0.0410	0.0250	0.0000
8	0.0280	0.1280	0.0150	0.0900	0.0540	0.0810	0.1630	0.0480	0.0000	0.0720	0.2500	0.0000	0.0000	0.0000	0.0720	0.0000
9	0.1240	0.3970	0.0430	0.0050	0.1240	0.0780	0.0530	0.0350	0.0220	0.0000	0.0270	0.0190	0.0010	0.0090	0.0060	0.0570
10	0.0710	0.3810	0.0300	0.0000	0.1250	0.1480	0.0000	0.0000	0.0000	0.1130	0.0000	0.0200	0.0110	0.0170	0.0670	0.0170
11	0.0100	0.2930	0.0070	0.0400	0.0170	0.0190	0.0000	0.0350	0.2300	0.0220	0.0630	0.0000	0.0590	0.0000	0.0140	0.1920
12	0.0910	0.1640	0.0180	0.0270	0.0190	0.0040	0.0000	0.0480	0.4110	0.0460	0.0790	0.0080	0.0000	0.0180	0.0510	0.0160
13	0.0130	0.0350	0.0080	0.0310	0.0090	0.0170	0.1280	0.0430	0.3040	0.0230	0.0540	0.0400	0.0850	0.0000	0.0360	0.1730
14	0.2020	0.0000	0.0380	0.0000	0.2510	0.2400	0.0000	0.0000	0.0000	0.2420	0.0000	0.0000	0.0000	0.0040	0.0000	0.0230
15	0.0190	0.0510	0.0020	0.0440	0.0240	0.0210	0.0050	0.0340	0.5120	0.0480	0.0230	0.0610	0.1220	0.0090	0.0270	0.0000

Table 3. The clustered attractions based on the probability tourist graph.

Cluster ID	Center of cluster	Members (clustering vertices)
1	Grand Central Terminal	Meatpacking District, Brooklyn Bridge, The Metropolitan Museum of Art, Statue of Liberty, Ground Zero Memorial, The Museum of Modern Art, SoHo, Grand Central Terminal
2	Times Square	Empire State Building, Times Square, Madison Square Garden, The High Line, World Trade Center, Radio City Music Hall, Top of The Rock, Central Park

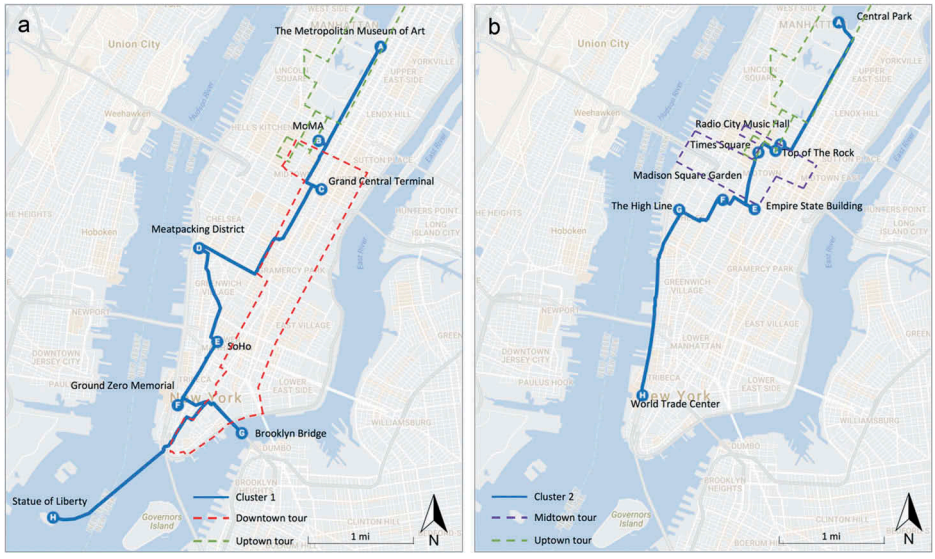

Figure 11. Geographic distribution of the clustered attractions and the recommended routes for tourists based on the Markov clustering result: (a) the first cluster; (b) the second cluster. The red dash line represents the route of Big Bus downtown tour; the green dash line represents the route of Big Bus uptown tour; the purple dash line represents the route of Big Bus midtown tour.

ORCID

Fei Hu http://orcid.org/0000-0001-5231-2303
Zhenlong Li http://orcid.org/0000-0002-8938-5466

References

Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008). Finding high-quality content in social media. *Proceedings of the 2008 International Conference on Web Search and Data Mining*, 183–194. doi:10.1145/1341531.1341557

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2), 163–177. doi:10.1080/0022250X.2001.9990249

Chen, S., Yuan, X., Wang, Z., Guo, C., Liang, J., Wang, Z., Zhang, X. L., & Zhang, J. (2016). Interactive visual discovering of movement patterns from sparsely sampled geo-tagged social media data. *IEEE Transactions on Visualization and Computer Graphics*, 22(1), 270–279. doi:10.1109/TVCG.2015.2467619

Cranshaw, J., Schwartz, R., Hong, J., & Sadeh, N. (2012). The livehoods project: Utilizing social media to understand the dynamics of a city. *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han & U. Fayyad (Eds.) Proceedings of the Second International Conference on Knowledge Discovery and Data Mining - KDD, (pp. 226–231) Palo Alto, CA: AAAI Press.

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, 69(4), 211–221. doi:10.1007/s10708-007-9111-y

Guo, D., & Chen, C. (2014). Detecting non-personal and spam users on geo-tagged Twitter network. *Transactions in GIS*, 18(3), 370–384. doi:10.1111/tgis.2014.18.issue-3

Guo, Q., & Karimi, H. A. (2017). A novel methodology for prediction of spatial-temporal activities using latent features. *Computers, Environment and Urban Systems*, 62, 74–85. doi:10.1016/j.compenvurbsys.2016.10.006

Haldrup, M. (2004). Laid-back mobilities: Second-home holidays in time and space. *Tourism Geographies*, 6(4), 434–454. doi:10.1080/1461668042000280228

Hasan, S., Zhan, X., & Ukkusuri, S. V. (2013). Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*. doi:10.1145/2505821.2505823

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), 97–109. doi:10.1093/biomet/57.1.97

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3), 260–271. doi:10.1080/15230406.2014.890072

Haythornthwaite, C. (2005). Social networks and Internet connectivity effects. *Information, Community & Society*, 8 (2), 125–147. doi:10.1080/13691180500146185

Hu, Y., Gao, S., Janowicz, K., Yu, B., Li, W., & Prasad, S. (2015). Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems*, 54, 240–254. doi:10.1016/j.compenvurbsys.2015.09.001

Huang, Q., Li, Z., Li, J., & Chang, C. (2016). Mining frequent trajectory patterns from online footprints. *Proceedings of the 7th ACM SIGSPATIAL International Workshop on GeoStreaming*. doi:10.1145/3003421.3003431

Huang, Q., & Wong, D. W. (2015). Modeling and visualizing regular human mobility patterns with uncertainty: An example using Twitter data. *Annals of the Association of American Geographers*, 105(6), 1179–1197. doi:10.1080/00045608.2015.1081120

Huang, Q., & Wong, D. W. (2016). Activity patterns, socio-economic status and urban spatial structure: What can social media data tell us? *International Journal of Geographical Information Science*, 30(9), 1873–1898. doi:10.1080/13658816.2016.1145225

Huang, X., Wang, C., & Li, Z. (2018a). A flooding probability reconstruction approach by enhancing near real-time imagery with real-time gauges and tweets. *IEEE Transactions on Geoscience and Remote Sensing*. doi:10.1109/TGRS.2018.2835306

Huang, X., Wang, C., & Li, Z. (2018b). A near real-time flood mapping approach by integrating post-event with satellite imagery and flood-related tweets. *Annals of GIS*, 24(2), 113–123. doi:10.1080/19475683.2018.1450787

Jankowski, P., Andrienko, N., Andrienko, G., & Kisilevich, S. (2010). Discovering landmark preferences and movement patterns from photo postings. *Transactions in GIS*, 14(6), 833–852. doi:10.1111/tgis.2010.14.issue-6

Jiang, Y., Li, Z., & Ye, X. (2018). Understanding demographic and socioeconomic biases of geotagged Twitter users at the county level. *Cartography and Geographic Information Science*, 1–15. doi:10.1080/15230406.2018.1434834

Kisilevich, S., Mansmann, F., & Keim, D. (2010). P-DBSCAN: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos. *Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application*. doi:10.1145/1823854.1823897

Kurashima, T., Iwata, T., Irie, G., & Fujimura, K.(2010). Travel route recommendation using geotags in photo sharing sites. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 579–588. doi:10.1145/1871437.1871513

Lau, G., & Mckercher, B. (2007). Understanding tourist movement patterns in a destination: A GIS approach. *Tourism Hosp Res* 7, 39–49. doi: 10.1057. palgrave. thr.

Lee, J. Y., & Tsou, M. H. (2018, January). Mapping spatio-temporal tourist behaviors and hotspots through location-based photo-sharing service (Flickr) data. In K. Peter, H. Huang, N. Weghe, & M. Raubal (Eds.), *LBS 2018: 14th International conference on location based services* (pp. 315–334). Cham, Switzerland: Springer.

Leung, X. Y., Wang, F., Wu, B., Bai, B., Stahura, K. A., & Xie, Z. (2012). A social network analysis of overseas tourist movement patterns in Beijing: The impact of the Olympic Games. *International Journal of Tourism Research*, 14(5), 469–484. doi:10.1002/jtr.v14.5

Lew, A. A., & Mckercher, B. (2002). Trip destinations, gateways and itineraries: The example of Hong Kong. *Tourism Management*, 23(6), 609–621. doi:10.1016/S0261-5177(02)00026-2

Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, 40(2), 61–77. doi:10.1080/15230406.2013.777139

Li, Z., Wang, C., Emrich, C. T., & Guo, D.(2017). A novel approach to leveraging social media for rapid flood mapping: A case study of the 2015 South Carolina floods. *Cartography and Geographic Information Science*, 45(2), 97–110, doi:10.1080/15230406.2016.1271356.

Lu, X., Wang, C., Yang, J. M., Pang, Y., & Zhang, L.(2010). Photo2trip: Generating travel routes from geo-tagged photos for trip planning. *Proceedings of the 18th ACM International Conference on Multimedia*, 143–152. doi:10.1145/1873951.1873972

Malik, M. M., Lamba, H., Nakos, C., & Pfeffer, J. (2015). Population bias in geotagged tweets. *People*, 1(3,759.710), 3,759.710–757,233.531.

Martin, Y., Li, Z., & Cutter, S. (in press). Leveraging Twitter to gauge evacuation compliance: Spatiotemporal analysis of Hurricane Matthew. *PLos ONE*. doi:10.1371/journal.pone.0181701

Mckercher, B., & Lau, G. (2008). Movement patterns of tourists within a destination. *Tourism Geographies*, 10(3), 355–374. doi:10.1080/14616688.2011.598542

Mckercher, B., & Lew, A. A. (2004). Tourist flows and the spatial distribution of tourists. In A. Lew, C. Hall, & A. Williams (Eds.), *A Companion to Tourism*, (pp. 36-48). Victoria: Blackwell Publishing.

Mckercher, B., Shoval, N., Ng, E., & Birenboim, A. (2012). First and repeat visitor behaviour: GPS tracking and GIS analysis in Hong Kong. *Tourism Geographies*, 14(1), 147–161. doi:10.1080/14616688.2011.598542

Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., … & Xin, D. (2016). Mllib: Machine learning in apache spark. *The Journal of Machine Learning Research*, 17(1), 1235–1241. http://jmlr.org/papers/v17/15-237.html

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's Firehose. In Proceedings of the Seventh International Conference on Weblogs and Social Media - *ICwSM* 2013 (pp. 400–408). Palo Alto, CA: AAAI Press. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/download/6071/6379

Noulas, A., Scellato, S., Mascolo, C., & Pontil, M. (2011). An empirical study of geographic user activity patterns in foursquare. In Proceedings of the Fifth International Conference on Weblogs and Social Media 2011 (pp. 570–573). Palo Alto, CA: AAAI Press.

NYC & Company (2018), Retrieved April 22, 2018 from https://business.nycgo.com/press-and-media/press-releases/articles/post/mayor-de-blasio-and-nyc-company-announce-nyc-welcomed-record-628-million-visitors-in-2017/,

Panteras, G., Wise, S., Lu, X., Croitoru, A., Crooks, A., & Stefanidis, A. (2015). Triangulating social multimedia content for event localization using Flickr and Twitter. *Transactions in GIS*, 19(5), 694–715. doi:10.1111/tgis.2015.19.issue-5

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: Real-time event detection by social sensors. *Proceedings of the 19th International Conference on World Wide Web*, 851–860. doi:10.1145/1772690.1772777

Sakaki, T., Okazaki, M., & Matsuo, Y. (2013). Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering*, 25(4), 919–931. doi:10.1109/TKDE.2012.29

Schneider, C. M., Belik, V., Couronné, T., Smoreda, Z., & González, M. C. (2013). Unravelling daily human mobility motifs. *Journal of the Royal Society Interface*, 10(84), 20130246. doi:10.1098/rsif.2013.0246

Shao, H., Zhang, Y., & Li, W. (2017). Extraction and analysis of city's tourism districts based on social media data. *Computers, Environment and Urban Systems*, 65, 66–78. doi:10.1016/j.compenvurbsys.2017.04.010

Stefanidis, A., Crooks, A., & Radzikowski, J. (2013). Harvesting ambient geospatial information from social media feeds. *GeoJournal*, 78(2), 319–338. doi:10.1007/s10708-011-9438-2

Sui, D. Z., Elwood, S., & Goodchild, M. (Eds.) (2012). *Crowdsourcing geographic knowledge: Volunteered geographic information (VGI) in theory and practice*. London: Springer doi:10.1007/978-94-007-4587-2

Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P., & Zhao, B. Y. (2009). User interactions in social networks and their implications. *Proceedings of the 4th ACM European Conference on Computer Systems*, 205–218. doi:10.1145/1519065.1519089

Yang, C., Huang, Q., Li, Z., Liu, K., & Hu, F. (2017). Big Data and cloud computing: Innovation opportunities and challenges. *International Journal of Digital Earth*, 10(1), 13–53. doi:10.1080/17538947.2016.1239771

Yang, C., Yu, M., Hu, F., Jiang, Y., & Li, Y. (2017). Utilizing cloud computing to address big geospatial data challenges. *Computers, Environment and Urban Systems*, 61, 120–128. doi:10.1016/j.compenvurbsys.2016.10.010

Yang, L., Wu, L., Liu, Y., & Kang, C. (2017). Quantifying tourist behavior patterns by travel motifs and geo-tagged photos from Flickr. *ISPRS International Journal of Geo-Information*, 6(11), 345. doi:10.3390/ijgi6110345

Zheng, Y.-T., Li, Y., Zha, Z. J., & Chua, T. S. (2011). Mining travel patterns from GPS-tagged photos. *International Conference on Multimedia Modeling*, 6523, 262–272.

Zheng, Y.-T., Zha, Z.-J., & Chua, T.-S. (2012). Mining travel patterns from geotagged photos. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3), 56.