World Wide Web (2019) 22:2655–2673
https://doi.org/10.1007/s11280-018-0551-8

# Understanding Skout users' mobility patterns on a global scale: a data-driven study

**Rong Xie[1,2,3] · Yang Chen[1,2,3] · Shihan Lin[1,2,3] · Tianyong Zhang[1,2,3] · Yu Xiao[4] · Xin Wang[1,2,3]**

Received: 18 December 2017 / Revised: 8 March 2018 / Accepted: 21 March 2018 /
Published online: 13 April 2018


**Abstract** Location-based social apps, such as Skout, have been widely used by millions of users for sharing their location information. In this work, we collected all the location information published by over 1.2 million Skout users during December 2012 and June 2016. Based on the collected information, we model the inter-city mobility of Skout users with a global city network, and analyze the evolution of the network based on its structural characteristics. Moreover, we look into Skout users' mobility patterns by discovering the most popular inter-city routes, destinations, and tightly connected city groups, and analyze the impact on the mobility patterns from geographical distances, languages and cultures.

This article belongs to the Topical Collection: *Special Issue on Web and Big Data*
Guest Editors: Junjie Yao, Bin Cui, Christian S. Jensen, and Zhe Zhao

✉ Yang Chen
chenyang@fudan.edu.cn

Rong Xie
xieronglucy@fudan.edu.cn

Shihan Lin
shlin15@fudan.edu.cn

Tianyong Zhang
tyzhang14@fudan.edu.cn

Yu Xiao
yu.xiao@aalto.fi

Xin Wang
xinw@fudan.edu.cn

[1] School of Computer Science, Fudan University, Shanghai, China

[2] Engineering Research Center of Cyber Security Auditing and Monitoring, Ministry of Education, Shanghai, China

[3] State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an, China

[4] Department of Communications and Networking, Aalto University, Espoo, Finland

Finally, we leverage machine learning techniques to build a model for identifying the most influential cities in the world according to the Skout data. The results are able to assist individuals, governors and business leaders in making better decisions regarding traveling, immigrating, measuring city improvements and cooperation with cities.

**Keywords** Human mobility · Skout · Global city network · PageRank

## 1 Introduction

Nowadays more and more people are moving across the world than ever before. Analyzing human mobility on a global scale is essential for business leaders, policy makers, as well as the general public. For example, it can help in improving the accuracy of location recommendations in the tourism industry [8, 14], in predicting the spread patterns of diseases [15], and in understanding the trends of resident migration [9].

There have been many data-driven human mobility studies [6, 10, 12, 26, 31, 32], whereas most of them focused on intra-city human mobility patterns. In other words, the analysis about inter-city human mobility on a global scale is still lacking. By analyzing human beings' mobility patterns between cities from all over the world, we are able to do a better job in many fields on a global scale such as city recommendations, prediction of disease spreading and understanding migration trend, which cannot be achieved by focusing on people's intra-city movements only. Therefore, research on inter-city human mobility patterns is of great importance.

In this paper we choose Skout,[1] a popular location-based social network (LBSN), as an example to study the user mobility on a global scale. We have collected the location information of over 1.2 million Skout users published during a period of 3.5 years. Based on the collected data, we describe the inter-city mobility with a *global city network*, analyze the patterns of user mobility based on the structural properties of the network, and study the influence of cities from the perspective of human mobility reflected by Skout data. Here a global city network can be considered as a variant of a place network [24], focusing on inter-city mobility instead of domestic mobility within a city. To the best of our knowledge, this is the first work that utilizes the location information collected from LBSNs to discover the patterns of inter-city human mobility on a global scale and study the city influence from the perspective of human mobility reflected by Skout data.

Our main contributions are summarized below:

Firstly, we model the inter-city mobility of Skout users on a global scale through a global city network. The network covers more than 240,000 inter-city trips between more than 18,000 cities in 184 countries.

Secondly, we analyze the structural characteristics of the global city network, and discover the user mobility patterns of Skout by identifying the most popular inter-city routes, destinations and city groups. Our key findings are listed below:

- Influential nodes in the structure of the global city network are major cities in the world. The influence of cities reflected by Skout data is consistent with that reflected by real world flight data.
- 99% of all the recorded inter-city trips of Skout users happen between cities that belong to a subset covering 56% of the city list.

[1] http://www.skout.com/

– Skout users prefer to travel to the cities nearby, or to the ones speaking the same languages or having similar cultures.

Thirdly, we design and implement a model to determine the level of a city's influence by using the Skout related features such as the number of buzzes created in the city. The F1-score is more than 0.8. It means that the mobility patterns shown in the location information of Skout are able to reflect the influence of cities in the world from a mobility perspective. More importantly, the knowledge of city influence level is beneficial to individuals, governors and business leaders in making better decisions regarding traveling, immigrating, measuring city improvements and cooperation with cities.

The rest of this paper is structured as below. We first discuss the related work in Section 2, and then introduce the method for collecting location information of Skout users in Section 3. Next, in Section 4, we conduct both static and dynamic analysis of the global city network built with the collected data. Then we analyze the Skout user mobility patterns and connections between cities in Section 5. Finally, we discover the ability of Skout data in reflecting the influence of cities in Section 6 before we conclude the work in Section 7.

## 2 Related work

Human mobility patterns have been studied in many works since they are of great importance and have a wide range of applications [8, 12, 15, 31]. Most of them used LBSNs' location data [9, 26, 31], cellular tower location recorded when phone calls were made [12], or GPS traces. We categorize the related work into two groups.

Characteristic analysis of human mobility. Cho et al. studied the location data collected from LBSNs and mobile phone users in a European country [9]. They found that people moved periodically in a confined region and tended to go to a distant place where there were friends there. Preoţiuc-Pietro and Cohn collected check-in data of more than 9000 Foursquare users and analyzed check-ins' temporal and spacial patterns [26]. Yu et al. collected nearly 12 million check-ins recording the mobility of more than 679,000 Fourquare users across 111 days [31]. They studied the spatial-temporal patterns of user activities and the prediction of users' transitions from one place to another by using spatial and temporal information in check-ins. Gonzalez et al. collected the location data of 100,000 randomly selected mobile phone users over a six-month period [12]. Each time a selected user initiated or received a phone call or a text message, her location would be recorded. Then they studied the characteristic of human trajectories and found that human trajectories were of high degree of temporal and spacial regularity. Brockmann et al. tried to figure out qualitative features of human mobility in the United States by observing the circulation of bank notes [5]. They found that human traveling behavior could be described accurately by a two-parameter continuous-time random walk model with scale-free jumps and long waiting times between displacements as well. Noulas et al. investigated the human mobility prediction by performing link prediction in the place network [24]. Differences between our work and these works are that we conduct an analysis of human trajectories on a global scale and on city granularity. Besides, we aim to find out the characteristics of human mobility on a global scale as well as the reasons that motivate people to move between different cities.

Applications and services based on human mobility analysis. Some works studied the check-in data[2] collected from LBSNs to infer users' preferences for different venues. Thus,

[2]A check-in of a user is a record containing the time and venue of her visit.

they are able to recommend venues for users. Bao et al. investigated each user's check-in history and inferred the user's preferences for venues. At the same time, they identified local experts who had more knowledge about a certain kind of venue type. By considering both individual user's preferences for venues and the opinions from local experts, they were able to recommend venues for each user. Their method had both high efficiency and accuracy [1]. Lian et al. conducted matrix factorization on check-in data and made the venue recommendation for users. By applying a finding on human mobility pattern by other researchers, i.e., individual visiting locations tended to cluster together, they optimized existed matrix factorization and gained better recommendation performance [20]. Other works inferred functionalities of different areas within a city from human mobility data. Çelikten et al. designed a model to identify the most distinctive features of a geographical area, utilizing the information of mobility data such as the location, time and corresponding activities. By the extraction of features of each geographical area, they were able to discover similar regions in different cities [7]. Cranshaw et al. introduced a clustering model and utilized the human mobility data to study the dynamics, structure, and characteristics of a city. They divided a city into different clusters (Livehoods) according to the types of places found in a certain area as well as people who appeared there [11]. Different from previous works, we aim to find the relationship between the human mobility and the influence of different cities in the global city network.

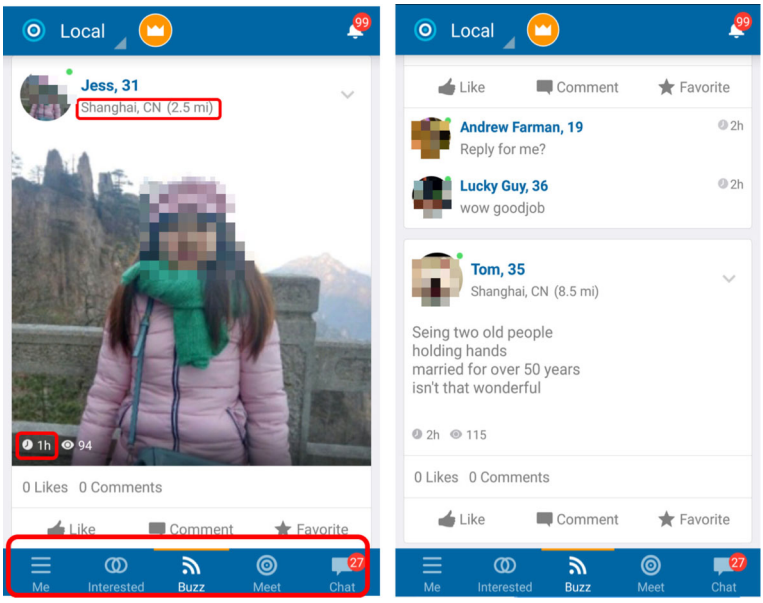
## 3 Data collection

Skout is an LBSN service for meeting new people. It obtains users' location information (i.e. the city where the user is staying) using GPS technology. By 2015, Skout had expanded its service to more than 100 countries and had over 10 million users.[3] In this work, we choose Skout as an example to study the inter-city mobility behavior of LBSN users.

Skout users can create and publish buzzes (a.k.a posts) using Skout mobile app. A buzz can be a picture, a piece of text, or their combination. Each buzz also contains meta-data, including the created location (with the granularity of city) and the created time. Two example buzzes are shown in Figure 1. Because all the Buzz pages can be accessed by any Skout user without restriction, we are allowed to collect all the public information on any Skout user's Buzz page. Based on the created location and time of each buzz, we are able to track each user's trajectories.

We ran a self-developed crawler from May 1st, 2016 to July 1st, 2016 to collect all the buzzes published by 1,220,560 Skout users. These Skout users are randomly picked from the whole set of over 10 million users. We collected in total 3,746,387 buzzes which were published between December 20th, 2012 and July 1st, 2016. These buzzes are later used for building the global city network. Note that Skout only allows us to download the 400 most recent buzzes of each user. Luckily, only 18 out of the 1,220,560 Skout users had over 400 buzzes. Therefore, it is safe to assume that our dataset includes the complete set of buzzes available on the randomly chosen Skout users' Buzz pages. What is more, Skout has a feature called "travel" which allows users to virtually set their location in any city. As the actual location may be different from the virtual one, we filter out the buzzes created when the user was using the "travel" feature. In practice, 6511 out of the total 3,746,387 buzzes (0.17%) were removed from our dataset.

[3] https://intransit.blogs.nytimes.com/2015/03/03/an-app-that-connects-travelers-with-locals/




(a) Buzz that includes a picture (b) Buzz in the form of text

**Figure 1** Screenshots of the Buzz page in Skout mobile app. In (**a**), the top red circle indicates the location where the buzz was created, the circle in the middle shows the length of time the buzz has been online, and the circle at the bottom indicates 5 categories of pages in Skout, i.e., Me, Interested, Buzz, Meet and Chat

## 4 Construction and analysis of the global city network

To gain a better understanding of Skout users' mobility patterns, we model the inter-city mobility with a global city network, and analyze the network's structural characteristics. In addition, we analyze the influence of cities based on our Skout dataset, and validate the results with real-world flight records.

### 4.1 Definition of the global city network

Noulas et al. [24] proposed the notion of "place network" to characterize the linkage between venues within a certain city. We propose to model the Skout user mobility using a global city network, a variant of the place network [24].

In this work we focus on the inter-city mobility of Skout users on a global scale. We define a global city network as a weighted directed graph $G^t$, with $t$ representing the time period during which the buzzes used to build the network were created. $G^t$ is composed of a node set $N^t$ and an edge set $E^t$. Nodes are cities in the world, while edges indicate direct transitions between cities within the time window $t$. A direct transition of a user between two cities $i$ and $j$ means that the user published a buzz in city $i$ first, followed by a buzz published in city $j$ within the next 36 hours. We limit the time window to 36 hours, assuming that a user moves directly from one city to another if she publishes buzzes in these two cities sequentially within 36 hours. The weight of an edge is the number of trips from one node to the other. However, there may still be transitions between cities we fail

to capture if a Skout user does not create buzzes in each of the cities she has visited. Nevertheless, given the willingness of LBSN users to publish their real-time locations, using the spatial-temporal data published in LBSNs to study human mobility is still a common practice [8, 23, 24].

### 4.2 Static analysis of the global city network

We construct the global city network based on all the location data we have collected. Structural properties of it are described in Table 1, from which we could obtain an overall understanding of the global city network's structure. The metrics mentioned in this section are explained below.

- **Nodes / Edges #:** The number of nodes (cities)/edges (direct transitions) in the network.
- **Weighted in-degree, weighted out-degree:** The accumulated weight of an edge connecting to or from the node in question. In other words, weighted in-degree and weighted out-degree of a node represent the number of times people travel to and from a city, respectively [2].
- **Density:** It equals to $\frac{m}{n(n-1)}$, where $m$ is the number of edges and $n$ is the number of nodes.
- **Clustering coefficient:** Clustering coefficient of a node is defined as the fraction of the number of edges between that node's neighbors and the maximum number of edges that could exist. Specifically speaking, in a directed graph, if a node has $N$ neighbors and there are $E$ edges between those neighbors, then the clustering coefficient of that node is calculated as $E/[N * (N - 1)]$. A high clustering coefficient means that the node tends to form tightly connected groups with their neighbors.
- **Shortest path:** It is a path between two nodes that has the least number of edges.
- **Diameter:** The length of the longest "shortest path" in a graph.
- **Assortativity:** The assortativity coefficient of a network reflects the tendency of a node in it to connect with other nodes with similar weighted degrees. It can be calculated as the Pearson correlation coefficient of weighted degrees of all linked node (city) pairs. Weighted degree of a node here means the sum of weights of edges connected to that node. The value of assortativity coefficient is between -1 and 1. If the value is positive then the network is assortative and otherwise the network is not assortative. The absolute value of the assortativity coefficient reflects that to what extend the network is assortative or not assortative.

**Table 1** Structural properties of the global city network

| Metric | The entire network | LSCC | LWCC |
|---|---|---|---|
| Nodes # | 18,444 | 6835 | 10,320 |
| Edges # | 51,819 | 46,921 | 51,382 |
| Density | 0.000152 | 0.001005 | 0.000482 |
| Assortativity | − 0.063 | − 0.068 | − 0.065 |
| Average weighted in-/out-degree | 6.5 | 16.8 | 11.62 |
| Average clustering coefficient | 0.098 | 0.257 | 0.175 |
| Average shortest path length | / | 3.9 | 4.1 |
| Diameter | / | 8 | 15 |

- **SCC and LSCC:** A strongly connected component (SCC) of a directed graph is a subgraph that all nodes are strongly connected. In other words, for any node pair $(u, v)$ in this subgraph, there is a directed path from $u$ to $v$, and a directed path from $v$ to $u$. Meanwhile, no additional edge or node can be added to this subgraph without breaking the property of being strongly connected. LSCC refers to the largest SCC.
- **WCC and LWCC:** If we convert all edges of a directed graph into undirected ones, a weakly connected component (WCC) can be defined if there is a path between any node pair in this subgraph, and no additional edge or node can be added to this subgraph without breaking the weakly connected property. LWCC refers to the largest WCC.

We first conduct basic measurements of the global city network. As shown in Table 1, there are 18,444 nodes and 51,819 edges in the global city network. The average clustering coefficient in the network is 0.098. It is smaller than that of the Facebook social network (0.164) but higher than that of the Renren social network (0.063) [30]. On the other hand, the average clustering coefficient of the LSCC and the LWCC in the global city network are 0.257 and 0.175, respectively. These values show higher level of local clustering than random graphs [30]. Furthermore, the shortest path length of the LSCC and the LWCC is on average 3.9 and 4.1, respectively. A high average clustering coefficient and low shortest path length indicate that the LSCC and the LWCC of the global city network are small-world networks [29].

Influential nodes in the structure of the global city network are major cities in the world. We define major cities as the cities satisfying one of the following criteria: a. having a population of over 1 million; b. being the political, economic or the cultural center of a country. c. being the transportation hub of a country. Statistics of human mobility between cities can reflect the influence of the cities from mobility perspective. To identify influential cities in the world, we calculate PageRank [25] for each node in the global city network. PageRank is a widely used metric that quantifies the influence of nodes within a network [17, 22, 28]. It has been introduced by the Google search engine to rank the websites. For any node of the network, its PageRank value is between 0 and 1. Since cities in the global city network are connected through Skout users' movements, PageRank values of cities in such network could reflect their influence from human mobility perspective. A higher PageRank value means that the corresponding node is more influential. A city is more influential in the global city network means that there are more people coming to that city from cities that are also quite influential.

Table 2 shows the nodes (cities) with top 10 PageRank values in the global city network. We find that those influential nodes (cities) having highest PageRank values are also major cities. Cities in Table 2 all have a population over 1 million and they feature in politics, economy, culture or traffic. New York City is the global financial center; Los Angeles is the major entertainment industry center in USA; Sydney is one of the most well-known cities in Australia; London, Jakarta, Seoul, Manila, Tokyo and Bangkok are all capital cities.

The influence of cities reflected by Skout data is consistent with that reflected by real-world flight data. To see whether the PageRank calculated using Skout data is consistent with the result generated by real-world mobility data, we compare cities' PageRank levels in two different networks. One is built from domestic USA flight records[4] and we call this network *F network*, while the other is built from the Skout data which we call *S network*. We use both flight and Skout data generated from January 1st, 2013 to December

[4] https://www.transtats.bts.gov/Tables.asp?DB_ID=120&DB_Name=Airline%20On-Time%20Performance%20Data&DB_Short_Name=On-Time

**Table 2** Nodes (cities) with top 10 PageRank. Influential nodes in the structure of the global city network are also major cities in the world

| City | PageRank |
|---|---|
| New York City | 0.0157 |
| London | 0.0124 |
| Los Angeles | 0.0104 |
| Jakarta | 0.0075 |
| Chicago | 0.0070 |
| Seoul | 0.0069 |
| Manila | 0.0055 |
| Tokyo | 0.0055 |
| Sydney | 0.0052 |
| Bangkok | 0.0050 |

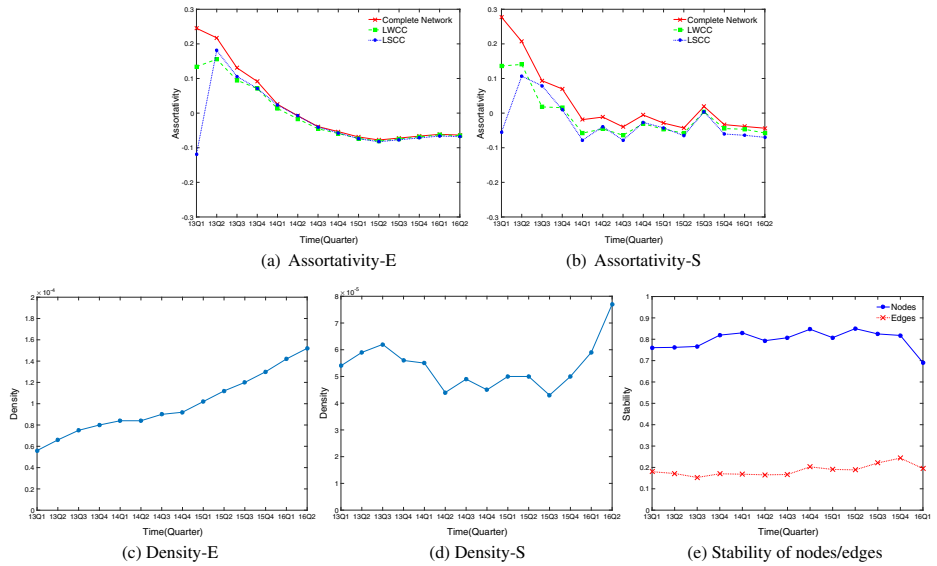
31st, 2015. Each flight record represents a flight and contains the following information: departure/arrival city, departure/arrival date, departure/arrival time. The reason why we use domestic flight data in USA to represent the human mobility in USA is that flights are very common transportation methods in USA so it can better represent inter-city movements of Americans. In the F network, a node is a USA city and a directed edge from one city to the other indicates that there has been a direct flight record between them. The nodes in S network are confined within USA, and the definition of edges is the same with global city network. After sorting the cities by PageRank value in each of the two city networks, we find 70% and 58% of cities with top 10 and top 50 PageRank values coincide, respectively. It means that the influence of cities reflected by Skout data is consistent with that reflected by real-world data. The reasons why we use Skout data instead of flight data to study the human mobility in our work are as follows. A Skout buzz reflects an individual's location while a record of flight data aggregates a group of people's movements. Also, Skout data contains not only locations but also texts, pictures and friends' comments. Thus, the Skout data is able to provide more information about people's movements and is beneficial for our study from the long run.

By analyzing the global city network from a static view, we find out its structural characteristics and the fact that Skout data is able to reflect the influence of cities in the world.

### 4.3 Dynamic analysis of the global city network

Global city network evolves over time. Therefore, looking at this network from a dynamic perspective helps us gain a deeper insight into it. In this subsection, we focus on the evolutionary aspect of the global city network.

To examine the evolution of the network, we take temporal factors into account and build two groups of global city networks. We use the location information in buzzes which were created from January 1st 2013 to June 30th 2016 covering 14 complete quarters. The first group contains 14 global city networks in which the $i$-th network is constructed based on the aggregated data of the first $i$ quarters, and we call them *evolutionary group*. The other group also contains 14 global city networks but the $i$-th network is constructed merely based on the data of the $i$-th quarter, and we call them *sequential group*. By observing the measurement results of the first group of city networks, we can see how the global city network grows, while from the second group, we can see the variation of the network in each quarter.



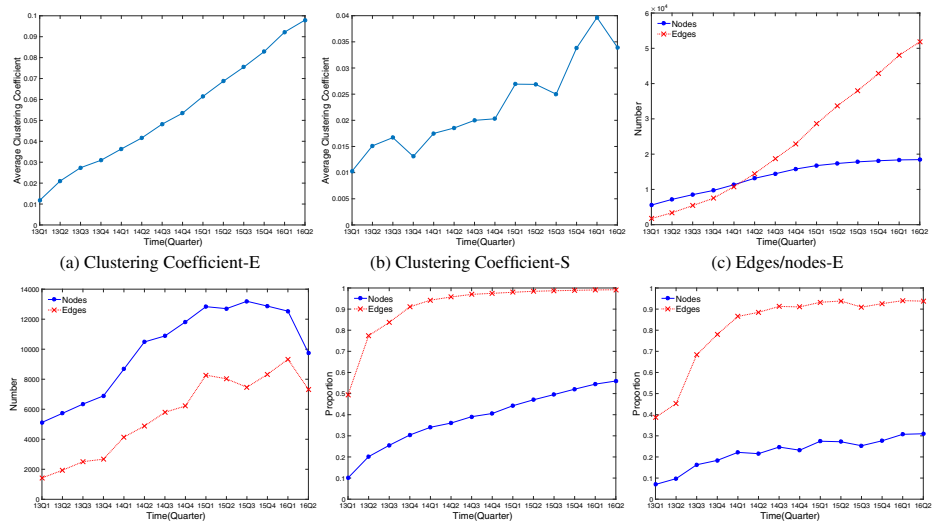

(a) Assortativity-E (b) Assortativity-S

(c) Density-E (d) Density-S (e) Stability of nodes/edges

**Figure 2** Assortativity, density, and stability of nodes/edges

The x-axis of Figures 2 and 3 represents the sequential number of the 14 quarters, and the letter E and S in the subtitles of each figure represent the evolutionary and sequential groups of global city networks, respectively.

People do not necessarily travel from one popular destination to another one. Figure 2b shows the assortativity of each global city network in the sequential group. All of the assortativity values are around 0 excluding those of the first 2 quarters. Furthermore, as shown in




(a) Clustering Coefficient-E (b) Clustering Coefficient-S (c) Edges/nodes-E

(d) Edges/nodes-S (e) Proportion-E (f) Proportion-S

**Figure 3** Measurements reflecting how the connectivity evolves over time

Figure 2a, the assortativity is decreasing from a relative high value and becomes lower and lower as time grows. Since the assortativity coefficient of a network reflects the tendency of a node in it to connect with other nodes with similar weighted degrees, it means that nodes in the city network have less and less tendency to link to nodes of similar weighted degrees. This is different from online social networks such as Facebook, i.e., the assortativity of Facebook's social network is 0.17 [30]. Weighted in/out degree of a node represents the number of times people moving into/out of a city. So we can equivalently say that whether people would move between two cities does not depend on the total number of movements to or from these two cities, respectively. It indicates that people do not necessarily travel from one popular destination to another one.

More and more movements are made by Skout users. Figure 2c shows the the density of the global city network at the end of each quarter. The values are calculated from the 14 global city networks in the evolutionary group. Figure 2d shows the corresponding values of the 14 global city networks in the sequential group. From both figures we can see that the graph density increases over time and more and more movements are made by Skout users.

Individuals do not tend to repeat their movement routes in successive quarters. We then investigate the variation pattern of edges and nodes by observing a metric called *stability*. We define the node/edge stability of a global city network in the sequential group as follows:

$$S_q = \frac{|X^q \cap X^{q+1}|}{|X^q|} \tag{1}$$

where $X^q$ represents the node/edge set in the $q$-th global city network of the sequential group. $S_q$ means the percentage of nodes/edges that appear in both the $q$-th and $(q+1)$-th global city networks to the total number of nodes/edges in the $q$-th global city network. As shown in Figure 2e, nodes tend to have much better stability than edges. The edges' stability values keep at around 0.2, meaning that there are only about 20% of them would appear again in the next quarter's global city network. This indicates that individuals do not tend to repeat their movement routes in successive quarters. While for nodes, 80% of the them would still appear in the global city network of the next quarter, indicating that there would always be movements among the same group of cities.

Figure 3 shows how the connectivity of the global city network changes over time. We have the following remarks.

The global city network is more and more connected. Figure 3a and b show the changes of average clustering coefficient within the 14 quarters by measuring global city networks in the evolutionary and sequential groups, respectively. Average clustering coefficient is increasing over time. It means that the global city network is more and more connected, and nodes form more and more tightly connected groups with their neighbors. Figure 3c and d show the growing tendency of the number of nodes and edges, respectively. We can see that there are more and more nodes and edges over time. Figure 3e and f show the growth of the proportion of the nodes and edges in the LWCC to that in the complete global city network (the LSCC has the same tendency so we did not show the figures for space saving). The larger the proportion is, the more connected the network is. Figure 3a–f together show that the global city network becomes better and better connected over time.

The growth of edges is much quicker than that of nodes. From Figure 3c we can see that the growth of edges is much quicker than that of nodes. It makes sense since there are fixed number of cities in the world but people never stop moving among cities. Each new movement will generate a new edge.

Nearly all Skout users' mobility is confined within fewer than 60% of all appeared cities. As shown in Figure 3e, until the last quarter, the edges in the LWCC occupy 99% of all

edges, while that of nodes is 56%, signifying that nearly all Skout users' mobility is confined within fewer than 60% of all appeared cities.

By observing the global city network from a dynamic view, we know how the global city network evolves and how Skout users' mobility patterns change along with time.

## 5 Human mobility patterns and connections between cities

In Section 4, we explore the structural characteristics of the global city network describing human mobility and connections between cities. In this section, we look deeper into mobility patterns and connections between cities by analyzing the global city network based on *global inter-city mobility map*. Furthermore, by clustering nodes in the global city network, we explore how cities in the world are divided into groups according to human mobility.

### 5.1 City pairs

We use a JavaScript library named D3.js[5] to visualize the city pairs with edges that have top 300 weights (there are 51,819 city pairs having edges between them in all). They are shown in Figure 4. We call it *global inter-city mobility map*. An edge is directed but for conciseness we do not show the arrow of the edge. Instead, we use curvature to show the direction of an edge: upward curvature means that the edge is pointed to the city on its right, and downward curvature means that the edge is pointed to the city on its left. For Figure 4, we have the following remarks.

First, there are more direct transitions between cities in the same country, such as USA, Indonesia, and Philippines. This is expectable, given the difference in distance, transport connectivity and traveling cost.

Second, cities that have many connections with other cities (they are easily noticed on the global inter-city mobility map) are those global or domestic transportation hubs. Examples include New York City, Chicago, Los Angeles and San Francisco in USA, London in UK, Sydney and Melbourne in Australia, Jakarta in Indonesia, and Manila in Philippines. This implies that there are more people moving into or out of transportation hubs. Moreover, there are stronger connections between big cities than between big cities and small cities, or, between small cities. Big cities here refer to those cities that have important roles in economy, politics, culture, or tourism, nationally or internationally.

Third, there are more connections between cities in the Northern Hemisphere than between Southern and Northern hemispheres. Those connections bridging Southern and Northern Hemisphere include cities in Chile and Spain, Argentina and Philippines, Britain and Australia. Flight duration between cities in these countries, such as flights from Santiago (Chile) to Madrid (Spain), from Buenos Aires (Argentina) to Manila (Philippines), is more than 12 hours, indicating a long distance. However, cities in these countries still have frequent human mobility. Chile, Argentina and Philippines all had been deeply impacted by Spain for hundreds of years from the 15th century to 19th century due to some historical reasons. Such kind of impact brought them cultural blending and immigration waves, and people in Spain, Chile and Argentina speak the same language, Spanish. Also, quite some Filipinos can speak Spanish. To some extent, we can say that people in these four countries have a similar cultural background. Though they are far from each other, strong connections

[5] https://d3js.org

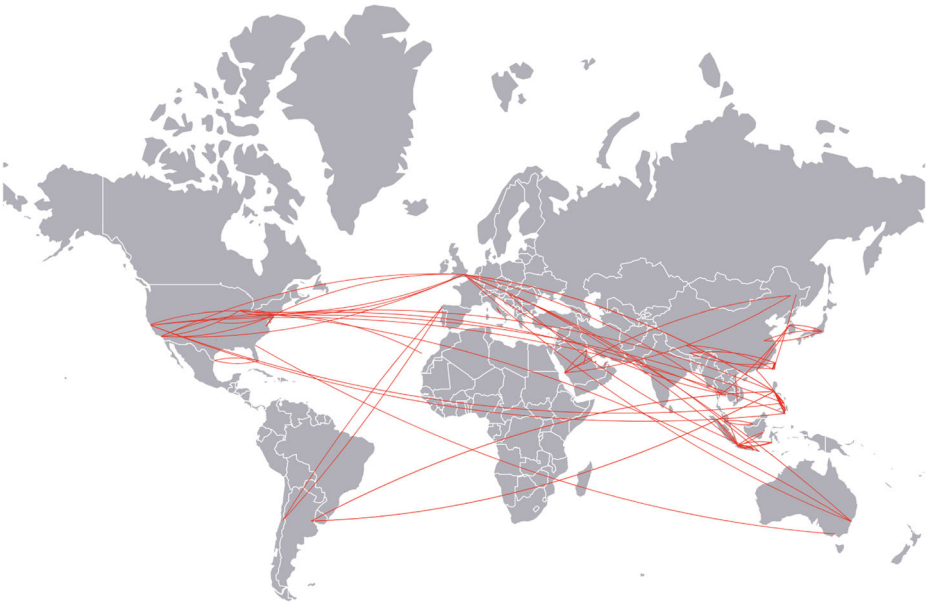


**Figure 4** Mobility map that shows the most popular inter-city routes and destinations. There are more direct transitions between cities in the same country; cities that have many connections with other cities are those transport hubs in the world or at least in their own country; there are more connections between cities in the Northern Hemisphere than that between Southern and Northern hemispheres

do exist between cities in these countries, indicating that language and culture have quite an amount of impact on human mobility. Frequent movements and strong connections between cities in Australia and UK, in USA and Philippines have the same implication.

Skour users prefer to move between cities close to each other, and between major cities in the world. We then sift out international city pairs in the global city network, i.e., city pairs having two cities in different countries. Table 3 shows international city pairs with top 10 edge weights between them. Since edges in the global city network are directed, city pairs are also directed. Each row in Table 3 represents an edge that is pointed from the

**Table 3** Ten most popular international inter-city routes. Skout users prefer to move between cities close to each other, and between major cities in the world

| Departure city | Arrival city |
|---|---|
| Singapore | Johor Bahru |
| Johor Bahru | Singapore |
| London | New York City |
| Singapore | Kampung Pasir Gudang Baru |
| New York City | London |
| London | Los Angeles |
| Bangkok | London |
| Kampung Pasir Gudang Baru | Singapore |
| Los Angeles | London |
| Bangkok | Seoul |

"Departure" city to the "Arrival" city. Kampung Pasir Gudang Baru and Johor Bahru are two Malaysian cities. There is a causeway called Johor Singapore Causeway linking the city of Johor Bahru and the town of Woodlands in Singapore, so it is convenient for people to move between these two places. What is more, Kampung Pasir Gudang Baru is located to the southeast of Johor Bahru and is only 32 kilometers away from it. These facts indicate that Singapore, Johor Bahru and Kampung Pasir Gudang Baru are close to one another, which we consider is one of the most important reasons that the edge weights between them are the highest ones. Other city pairs show that movements between significant international cities like New York City, London, Bangkok, and Seoul are frequent.

### 5.2 City groups

Besides connections between two cities, there are also strong connections within a group of cities.

We leverage a community detection method to find those cities having tight inter-city connections. We transfer the complete global city network to an undirected network by combining directed edges between the same pair of nodes into one undirected edge. We use the sum of their weights as the new weight. A community consists of a group of nodes that are linked to each other more densely than to the rest of the nodes. Equivalently, connections between different communities are sparser. Among many community detection methods, we choose Louvain algorithm [3] due to its high efficiency. To evaluate how well a network can be clustered into communities, the metric called modularity is widely used. What modularity measures is that the difference between the actual fraction of edges in communities and the fraction of edges in communities if edges are randomly connected. Its value is between $-1$ and 1. The larger the value is the better the network is clustered into communities. Modularity is defined in (2) [33]:

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j) \tag{2}$$

where $\delta(c_i, c_j)$ is 1 when node i and j are in the same community, otherwise it is 0. $k_i$ is the weighted degree of node i. $m$ is the total weights of all edges in the network. $A$ is the adjacency matrix of the network. $A_{ij}$ equals to the weight of the edge linking node i and j.

Modularity no less than 0.3 means that the corresponding network has been clustered well into communities [16]. We do community detection on our global city network (undirected version) consisting of 18,444 nodes and 51,819 edges, and its modularity is 0.733, much larger than 0.3. This means that the network has been clustered into communities well. Figure 5 visualizes the communities having top 8 sizes (excluding the largest one). Each dot represents a city, and cities in the same community are in the same color. Table 4 shows the country distribution in top 8 communities (excluding the largest one). The reason why we do not show the largest community is that this community actually covers the cities which cannot be allocated into any other city community. As a result, cities in the largest community do not satisfy our definition of city groups since they do not have strong connections to each other. As shown in Figure 5 and Table 4, city distribution in communities have the following characteristics.

First, cities in the same country have tighter connections. Cities in the same country are often in the same community, showing that cities in the same country have tighter connections.

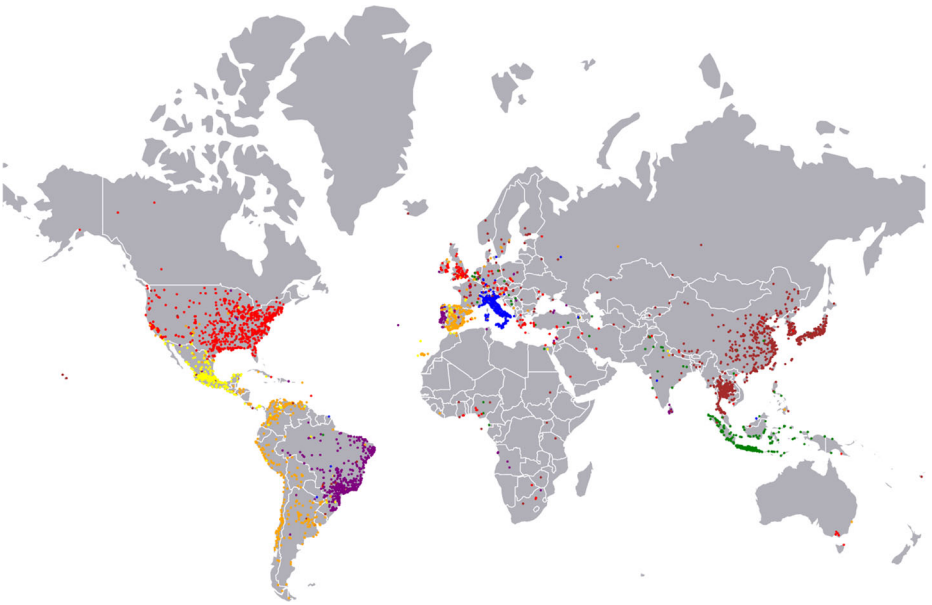


**Figure 5** Communities identified from the global city network. Each dot represents a city, and cities in the same community are in the same color. Cities in the same country have tighter connections; cities having the same language or same cultural background tend to be in the same community; many cities in China, Japan, South Korea and Thailand belong to the same community

Second, cities having the same language or same cultural background tend to belong to the same community, which supports our findings in the "city pairs" part. Most of the red dots represent cities in USA, UK and Australia. All of them are English speaking countries, and both USA and Australia were deeply impacted by UK in the past. A large number of UK people moved from UK to USA or Australia during the time of colonization. Large portion of orange dots fall in the area of Spain and the western area in South America; most purple dots appear in Brazil and Portugal. The orange and purple groups of cities have the same situation with the aforementioned UK, USA and Australia, having the same language and deep culture blending due to some historical reasons. Cities in the red, orange and purple groups respectively, all experienced deep culture blending due to historical reasons such as

**Table 4** Top 3 countries in top 8 communities

| Community rank | 1st country | 2nd country | 3rd country |
|---|---|---|---|
| 1 | USA (87.0%) | UK (3.1%) | Greece (1.7%) |
| 2 | Japan (30.4%) | China (29.2%) | Thailand (19.2%) |
| 3 | Spain (35.2%) | Argentina (12.2%) | Colombia (11.8%) |
| 4 | Brazil (81.4%) | Portugal (11.0%) | Sri Lanka (2.4%) |
| 5 | Italy (92.5%) | Switzerland (1.4%) | France (1.2%) |
| 6 | Indonesia (90.3%) | India (1.8%) | Belgium (0.8%) |
| 7 | Mexico (83.8%) | Panama (3.2%) | Spain (3.2%) |
| 8 | Philippines (88.8%) | USA (2.0%) | India (1.1%) |

colonization and speaking the same language, indicating that cities with these two features are more likely to have strong connections.

Third, many cities in China, Japan, South Korea and Thailand belong to the same community (marked by brown dots). These four countries are close to each other in east/southeast Asia, which makes it easier and more convenient for people moving from one to another. Another point is that China, Japan and South Korea had much cultural communications in the past, such as in Tang Dynasty (A.D. 618-A.D.907), culture in these three countries have some similarities and people there are familiar with each other's culture. This indicates that cities having shorter distances between them or similar culture are more likely to have frequent human mobility and strong connections.

### 5.3 Roles of cities

All the cities on the list of popular origins and destinations are big and important cities worldwide or at least domestically. We also inspect into the most popular move-in and move-out cities. The popularity of move in and out can be represented by the weighted in and out degrees of each city in the global city network. Table 5 lists the popularity rank. All the cities presented in the table appear in both the lists of top 10 most popular move in and move out cities, excluding Seoul, Bangkok and Taipei which are only in the list of top 10 move in or that of the top 10 move out cities. All the cities on the list are big and important cities in the world or at least in their own countries.

## 6 Influence of cities reflected by Skout data

In Section 4, we explained that PageRank could reflect the influence of cities from a mobility perspective, and that a city is more influential in the global city network means that there are more people coming to that city from cities that are also quite influential. Finding out the influence level of a city is beneficial to individuals, governors and business leaders in making better decisions regarding traveling, immigrating, measuring city improvements and cooperation with cities. For individuals, they are able to choose more influential cities to travel or immigrate; for governors, they could use the influence level of a city as a metric reflecting the city's importance and popularity from the perspective of human mobility; for business leaders, they could make better decisions in whether or not doing business in a city

**Table 5** The most popular origins and destinations. All the cities on this list are big and important cities in the world or at least in their own countries

| Rank | Move in | Move out |
|---|---|---|
| 1 | Jakarta | Jakarta |
| 2 | New York City | Manila |
| 3 | London | Port Area |
| 4 | Manila | Bangkok |
| 5 | Bagong Pagasa | Bagong Pagasa |
| 6 | Los Angeles | Mandaluyong City |
| 7 | Mandaluyong City | Taipei |
| 8 | Port Area | New York City |
| 9 | Seoul | London |
| 10 | Bekasi | Bekasi |

so that they are able to take the advantage of the influence of the selected city. For example, when a business leader aims to propagate a piece of information widely, besides the online propagation, she could propagate it offline. Then selecting more influential cities from the human mobility perspective to put the advertisement of the information is essential since the budget is limited and the propagation efficiency is critical. Therefore, the identification of influential cities in the global city network is of great importance and meaningfulness.

Unfortunately, calculating the PageRank value of a certain city is complicated because it requires the knowledge of the entire city network, i.e., the complete movement records reflected by buzzes of all the Skout users. For a certain city, if we are able to identify the level of a city's PageRank accurately with the Skout data of users from this city only, we are able to sense the influence of the city more efficiently.

### 6.1 Identifying influential cities in the F network

In Section 4, we compare the levels of cities' PageRanks in F and S network. We find out that Skout data can reflect the influence of cities according to human beings' movements in USA very well. As a result, we check if the features extracted from Skout data can perform well in identifying cities' PageRank levels in the F network. We do this work with supervised machine learning algorithms.

One hundred ninety-five cities appear in both of the two networks (F and S networks). We extract the following three features of each of the 195 cities from Skout data: *the number of buzzes, number of Skout users, average number of buzzes created by each user*. We also calculate the PageRank of each city in the F network and choose the PageRank value at the top 30% point as a threshold (there is a big gap between the top 30% PageRank values and the rest of them). Cities having PageRank values higher than the threshold are labeled as "1", meaning that they are more influential. While the rest are labeled as "0". To balance the training set, we randomly select similar number of samples labeled as "0" to the number of the samples labeled as "1".

We use Weka [13] to implement several classic machine learning algorithms, such as Random Forest [4], Logistic Model Trees (LMT) [18], Decision Tree (J48) [27] and Logistic Regression [19]. We use precision, recall and F1-score to quantify the identification performance. Precision is the fraction of cities identified to be more influential that really are more influential. Recall is the fraction of more influential cities that have been successfully identified. F1-score is the harmonic mean of precision and recall. In Weka, we use 10-fold cross-validation for evaluation. For each machine learning algorithm, we adjust parameters to pursue the highest F1-score. Table 6 shows the F1-score of four machine learning algorithms (in Weka). Random Forest algorithm performs the best and the average F1-score is 0.838. Therefore, we can conclude that cities' features extracted from Skout data perform

**Table 6** Identify PageRank level of cities in USA city network (F network)

| Algorithm | Parameter | Precision | Recall | F1-score |
|---|---|---|---|---|
| Random forest | 100 trees, 1 feature/tree | 0.848 | 0.839 | 0.838 |
| C4.5 (J48) | Confidence factor C = 0.25, Instance/leaf M = 2 | 0.717 | 0.715 | 0.714 |
| LMT | min instances = 20, boosting iteration = −1, beta value = 0.0 | 0.708 | 0.705 | 0.703 |
| Logistic regression | ridge = 1.0e-8 | 0.691 | 0.684 | 0.681 |

**Table 7** Identify PageRank level of cities in global city network (S netowrk)

| Algorithm | Parameter | Precision | Recall | F1-score |
|---|---|---|---|---|
| LMT | min instances = 15, boosting iteration = −1, beta value = 0.0 | 0.973 | 0.973 | 0.973 |
| C4.5 (J48) | Confidence factor C = 0.2, Instance/leaf M = 2 | 0.973 | 0.973 | 0.973 |
| Random forest | 100 trees, 1 feature/tree, max depth = 3 | 0.970 | 0.970 | 0.970 |
| Logistic regression | ridge = 1.0e-8 | 0.970 | 0.970 | 0.970 |

well in identifying levels of cities' PageRank in F network. Therefore, it is practical to use Skout data as an indicator for the status of a USA city, to see whether it is playing an influential role among all the cities.

### 6.2 Identifying influential cities in the global city network

Next, we identify PageRank level of each city in the global city network using features extracted from Skout data. We use the same features with the previous subsection. There are 18,444 cities in all. By observing their PageRank values, we find that there is a big gap between the top 1% PageRank values and the rest of them. Therefore, we choose the PageRank value at the top 1% point as a threshold. Cities having PageRank values higher than the threshold are labeled as "1", meaning that they are more influential. While the rest are labeled as "0". Also, we balance the number of samples in each class before we train the data. The average F1-score of two classes of four machine learning algorithms (in Weka) are listed in Table 7. LMT and C4.5 perform the best and the F1-score is 0.973. It means that Skout data can be used for determining which city has the top level of PageRank in global city network. Therefore, we can use Skout data in identifying which cities are playing influential roles in the world.

To sum up, the above model we design and implement is able to identify more influential cities in both of the F network constructed by real-world traffic data and the global city network constructed by LBSN location data in an accurate and efficient way.

## 7 Conclusion

In this paper, we conduct a data-driven study on Skout user mobility patterns and connections between cities according to the human mobility on a global scale. We first examine the structural characteristics of the global city network constructed from the location information of more than 1.2 million Skout users. We observe the network from both static and dynamic perspectives. We find out that influential nodes in the structure of the global city network are also major cities in the world and that the influence of cities reflected by Skout data is consistent with that reflected by real-world flight data. Next, we extract and visualize city pairs and city groups having strong connections, and find out that human mobility among cities show quite evident patterns. Geographical distance, language and cultures all have impact on human mobility patterns. Finally, we leverage machine learning techniques to build a model to determine the level of a city's influence by using the Skout related features such as the number of buzzes created in the city. The F1-score is more than 0.8.

Therefore, the mobility patterns shown in the location information of Skout buzzes reflect the influences of cities in the world from the mobility perspective.

In the future, in addition to the temporal and spatial data of Skout buzzes, we would also like to involve other kinds of information such as the text or images of buzzes into the analysis. By doing so, we are able to obtain a better understanding of the motivations behind the human beings' movements between cities and understand the influence of cities from more specific mobility perspectives. For example, we could group movements according to different motivations and investigate the influence of cities from each kind of mobility motivation. What is more, we could also involve temporal and spatial data collected from other LBSNs such as the Swarm app [21] to supplement the the Skout data.

**Acknowledgements** This work is sponsored by National Natural Science Foundation of China (No. 61602122, No. 71731004), Natural Science Foundation of Shanghai (No. 16ZR1402200), Shanghai Pujiang Program (No. 16PJ1400700), Academy of Finland (No. 268096).

## References

1. Bao, J., Zheng, Y., Mokbel, M.F.: Location-based and preference-aware recommendation using sparse geo-social networking data. In: Proceedings of ACM SIGSPATIAL (2012)
2. Barrat, A., Barthelemy, M., Pastor-Satorras, R., Vespignani, A.: The architecture of complex weighted networks. PNAS **101**(11), 3747–3752 (2004)
3. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech: Theory Exp. **2008**(10), P10,008 (2008)
4. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
5. Brockmann, D., Hufnagel, L., Geisel, T.: The scaling laws of human travel. Nature **439**(7075), 462–465 (2006)
6. Canzian, L., Musolesi, M.: Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In: Proceedings of ACM Ubicomp (2015)
7. Çelikten, E., Falher, G.L., Mathioudakis, M.: Modeling urban behavior by mining geotagged social data. IEEE Transactions on Big Data **3**(2), 220–233 (2017)
8. Cheng, C., Yang, H., Lyu, M.R., King, I.: Where you like to go next: successive point-of-interest recommendation. In: Proceedings of IJCAI (2013)
9. Cho, E., Myers, S.A., Leskovec, J.: Friendship and mobility: user movement in location-based social networks. In: Proceedings of ACM KDD (2011)
10. Cici, B., Gjoka, M., Markopoulou, A., Butts, C.T.: On the decomposition of cell phone activity patterns and their connection with urban ecology. In: Proceedings of ACM Mobihoc (2015)
11. Cranshaw, J., Schwartz, R., Hong, J.I., Sadeh, N.: The livehoods project: utilizing social media to understand the dynamics of a city. In: Proceedings of AAAI (2012)
12. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. Nature **453**(7196), 779–782 (2008)
13. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)
14. Hao, Q., Cai, R., Wang, C., Xiao, R., Yang, J.M., Pang, Y., Zhang, L.: Equip tourists with knowledge mined from travelogues. In: Proceedings of WWW (2010)
15. Hufnagel, L., Brockmann, D., Geisel, T.: Forecast and control of epidemics in a globalized world. PNAS **101**(42), 15,124–15,129 (2004)
16. Kwak, H., Choi, Y., Eom, Y.H., Jeong, H., Moon, S.: Mining communities in networks: a solution for consistency and its evaluation. In: Proceedings of ACM IMC (2009)
17. Kwak, H., Lee, C., Park, H., Moon, S.: What is twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web (2010)
18. Landwehr, N., Hall, M., Frank, E.: Logistic model trees. Mach. Learn. **95**(1-2), 161–205 (2005)
19. le Cessie, S., van Houwelingen, J.: Ridge estimators in logistic regression. Appl. Stat. **41**(1), 191–201 (1992)
20. Lian, D., Zhao, C., Xie, X., Sun, G., Chen, E., Rui, Y.: Geomf: joint geographical modeling and matrix factorization for point-of-interest recommendation. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2014)

21. Lin, S., Xie, R., Xie, Q., Zhao, H., Chen, Y.: Understanding user activity patterns of the swarm app: a data-driven study. In: ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2017 ACM International Symposium on Wearable Computers (2017)
22. Liu, Q., Xiang, B., Yuan, N.J., Chen, E., Xiong, H., Zheng, Y., Yang, Y.: An influence propagation view of pagerank. ACM Trans. Knowl. Discov Data **11**, 30:1–30:30 (2017)
23. Noulas, A., Scellato, S., Mascolo, C., Pontil, M.: An empirical study of geographic user activity patterns in foursquare. ICWSM 570–573 (2011)
24. Noulas, A., Shaw, B., Lambiotte, R., Mascolo, C.: Topological properties and temporal dynamics of place networks in urban environments. In: Proceedings of WWW (2015)
25. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. Stanford InfoLab Technical Report 1999–66 (1999)
26. Preoţiuc-Pietro, D., Cohn, T.: Mining user behaviours: a study of check-in patterns in location based social networks. In: Proceedings of ACM Websci (2013)
27. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco (1993)
28. Tang, J., Lou, T., Kleinberg, J.: Inferring social ties across heterogenous networks. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (2012)
29. Watts, D.J.: Networks, dynamics, and the small-world phenomenon. Am. J. Sociol. **105**(2), 493–527 (1999)
30. Wilson, C., Boe, B., Sala, A., Puttaswamy, K.P., Zhao, B.Y.: User interactions in social networks and their implications. In: Proceedings of ACM Eurosys (2009)
31. Yu, Y., Tang, S., Zimmermann, R., Aizawa, K.: Empirical observation of user activities: check-ins, venue photos and tips in foursquare. In: Proceedings of the 1st International Workshop on Internet-Scale Multimedia Management (2014)
32. Yuan, N.J., Zhang, F., Lian, D., Zheng, K., Yu, S., Xie, X.: We know how you live: exploring the spectrum of urban lifestyles. In: Proceedings of ACM COSN (2013)
33. Zhao, X., Sala, A., Wilson, C., Wang, X., Gaito, S., Zheng, H., Zhao, B.Y.: Multi-scale dynamics in a massive online social network. In: Proceedings of ACM IMC (2012)