Understanding Urban Human Activity and Mobility Patterns Using Large-scale Location-based Data from Online Social Media

Samiul Hasan
Research Assistant
School of Civil Engineering,
Purdue University
West Lafayette, IN 47907, US
samiul.hasan@gmail.com

Xianyuan Zhan
Research Assistant
School of Civil Engineering,
Purdue University
West Lafayette, IN 47907, US
zhanxianyuan@gmail.com

Satish V. Ukkusuri
Associate Professor
School of Civil Engineering,
Purdue University
West Lafayette, IN 47907, US
sukkusur@ecn.purdue.edu

ABSTRACT

Location-based check-in services enable individuals to share their activity-related choices providing a new source of human activity data for researchers. In this paper urban human mobility and activity patterns are analyzed using location-based data collected from social media applications (e.g. Foursquare and Twitter). We first characterize aggregate activity patterns by finding the distributions of different activity categories over a city geography and thus determine the purpose-specific activity distribution maps. We then characterize individual activity patterns by finding the timing distribution of visiting different places depending on activity category. We also explore the frequency of visiting a place with respect to the rank of the place in individual's visitation records and show interesting match with the results from other studies based on mobile phone data.

Keywords

Social media, large-scale, location-based data, human mobility pattern, urban activity pattern

1. INTRODUCTION

The introduction of location-based services in social media applications of smartphones has enabled people to share their activity related choices (check-in) in their virtual social networks (e.g. Facebook, Foursquare, Twitter etc.) providing unprecedented amount of user-generated data on human movement and activity participation. This data contains detailed geo-location information, which reflects extensive knowledge about human movement behavior. In addition, the venue category information for each check-in is recorded from which user activities can be inferred. Thus location-based data offers us a new dimension of information related to human activity categories with greater details. Researchers are realizing the potential to harness the rich information provided by the location-based data which has already enabled many novel applications such as recommendation system for physical locations (or activity) [1, 2] or recommending potential customers or friend [3, 4]; and determining popular travel routes in a city [5]. This data has the potential to impact many other areas including travel demand modeling, ubiquitous computing, epidemiology, urban planning, security and health monitoring. As such, a tremendous opportunity exists to develop fundamental tools to analyze this very large-scale spatial and temporal data that allows one to understand the social and behavioral characteristics of the users of location-based services.

Previous research efforts on individual activity-travel patterns over longer time periods were usually based on people's movements through traditional surveys on travel journeys [6, 7, 8]. The obtained information was based on questionnaires that are usually costly to implement and with intrinsic limitations to cover large number of individuals and some problems of reliability. These efforts, however, demonstrated that individual mobility patterns are strongly related with land-use patterns as well as the built environment of a city, and individual daily travel patterns exhibit great regularity [9, 10, 11, 12, 13].

On the other hand, there are some recent mobility studies that have used distance-based measures to characterize human mobility patterns using alternative datasets collected from mobile phones [14], bank notes movements [15], and subway smart-card transactions [16, 17] etc. These studies however limit the understanding of the interplay between selection of destinations for different activity purposes and mobility dynamics due to the lack of information about the purposes behind these movements. In this context, location-based data has received increasing attention in the research community, as the rich information in the data connects each geo-location record with a venue category indicating the purpose of the activity participated. In more recent studies, Cheng et al. [18] investigated 22 million check-ins and observed similar mobility pattern found in previous researches [14, 15], which is a mixture of short, random movements with occasional long jumps. Cho et al. [19] investigated the relationship between human mobility and social relationship using data from Gowalla and Brightkite. They found that social relationships can explain 10% to 30% of all human movements, while periodic behavior explains 50% to 70%. However, the dimension of human activity was not consid-

UrbComp '13, August 11-14, 2013, Chicago, Illinois, USA.
Copyright © 2013 ACM 978-1-4503-2331-4/13/08 ...$15.00.

ered in both of the researches. In this paper, by considering the temporal dimension and activity categories (i.e. purposes) into the analysis, we discover more realistic and detailed descriptions of human mobility dynamics. Considering the activity purposes in the analysis will enable us to develop advanced models for predicting mobility decisions.

We consider the location-based data obtained from online social media check-in services to characterize urban human activity and mobility patterns. We first investigate the characterization and visualization of aggregate human mobility and activity patterns by constructing a virtual grid reference of a city map into square cells of 200 by 200 meters. We discover a relationship between the popularity of a cell and the probability of visiting the cell. Spatial distributions of visiting different places are also determined for various activity purposes by counting the number of purpose-specific visits within each cell and computing the proportion of visits to each cell for each activity category. This generates activity distribution maps showing the popular places within a city and the functionality of each part of the urban area. Check-in distributions appear differently for different activity categories suggesting a strong influence of urban context on people's destination choices. Using Kernel density estimation methods we construct time-dependent activity density maps. Using this approach, we can also visualize different human activities in a city and thus capture the pulse of urban human activities.

Next we investigate the characterization of the spatio-temporal aspects of individual mobility patterns. We determine a set of statistical properties to characterize human mobility based on check-in data from online social media. First, we observe the timing of visiting different places depending on activity category. Second, we explore the frequency of visiting a place with respect to the rank of the place in individual's visitation records. Recently it has been suggested that the visitation frequency of the Lth most visited location is well approximated by Zipf's law: $P(L) \sim L^{-\eta}$, with $\eta \approx 1.2$ independent of N the total number of visited locations [14, 20].

In following sections we present the description of the data set and the findings related to individual mobility and urban activity patterns.

2. DATA COLLECTION

2.1 Dataset

The dataset used in this analysis is collected from a widely used social media tool called Twitter where users can post short messages up to 140 characters. These short messages are generally called status message in the social media norm and specifically called "Tweets" in Twitter. When permissions are given by the users, each of their tweets are attached with a corresponding geo-location. In addition to posting status messages, Twitter allows its users to post statuses from third-party "check-in" services (e.g. Foursquare). When Foursquare users "check-in" to a place this status can be posted to their Twitter pages. In this work we use a large-scale check-in data available from [18]. The dataset contains check-ins from Feb, 25, 2010 to January, 20, 2011. On average each user has 25 check-ins.

An example of a tweet with a "check-in" looks like:
tweet(79132591248261120)={189872633, ####,
79132591248261120, Fri Jun 10 10:27:34 +0000 2011,
40.7529422,-73.9780177, "I'm at Central Cafe & Deli
(16 Vanderbilt Ave., New York) http://4sq.com/jMS87x"}

Table 1: New York Dataset Details

Original dataset	
Number of users	20606
Number of check-ins	680564
Study sample	
Number of users	3256
Number of check-ins	504000

Table 2: Activity Category Classification

Activity Category	Type of Visited Location
Home	Home (private), Residential Building (Apartment/Condo)
Work	Office, Coworking Space, Tech Startup, Design Studio
Eating	Coffee Shop, Restaurant, Pizza, Burger, Caf, Diner, Steakhouse, Sandwich, Bakery, Breakfast, Bagel Shop, Taco Place, Gourmet Shop, Tea Room, etc.
Entertainment	Pub, Nightclub, Bar, Entertainment, Arcade, Theater, Club, Concert Hall, Other Nightlife, Dance Studio, Opera House, Casino, Event Space, etc.
Recreation	Park, Gym, Playground, Dog Run, Scenic Lookout, Beach, Lake, Zoo or Aquarium, Field, Tennis Court, Resort, Ski Area, Soccer Field, etc.
Shopping	Supermarket, Store, Plaza, Pharmacy, Bookstore, Mall, Farmers Market, Boutique, Miscellaneous Shop, Automotive Shop, Food & Drink Shop, etc.

After collecting the original dataset we select subsets of all the observations within three different cities in US, which are New York, Chicago and Los Angeles. We create a boundary region for these cities and extract all the check-in observations within that region. The New York dataset has largest amount of data, with 20606 users and 680564 check-in observations, while Chicago dataset has 7136 users and 193825 check-in observations, and Los Angeles dataset has 11298 users and 314783 check-in observations. We select New York dataset as our main dataset, and perform most of our analysis based on the New York dataset. However to find individual longitudinal mobility patterns we study only those users who have more than 50 check-ins. Some basic information about the New York dataset are given in Table 1. Chicago and Los Angeles dataset are used to conduct city-level comparison of popular places for different activities, which will be introduced in section 3.1.

2.2 Identification of Activity Categories

One of the major advantages of using location-based social media data is the ability to identify activity purposes. Each check-in observation reports a short link to the original location-based service provider (e.g. Foursquare). When queried in the location-based service provider, this link gives information about the category of the visited venue. We classify different activity categories based on the type of the visited locations (see Table 2). About 94.5% of the check-ins have any category information available; for rest of the check-ins their respective categories were not resolved.

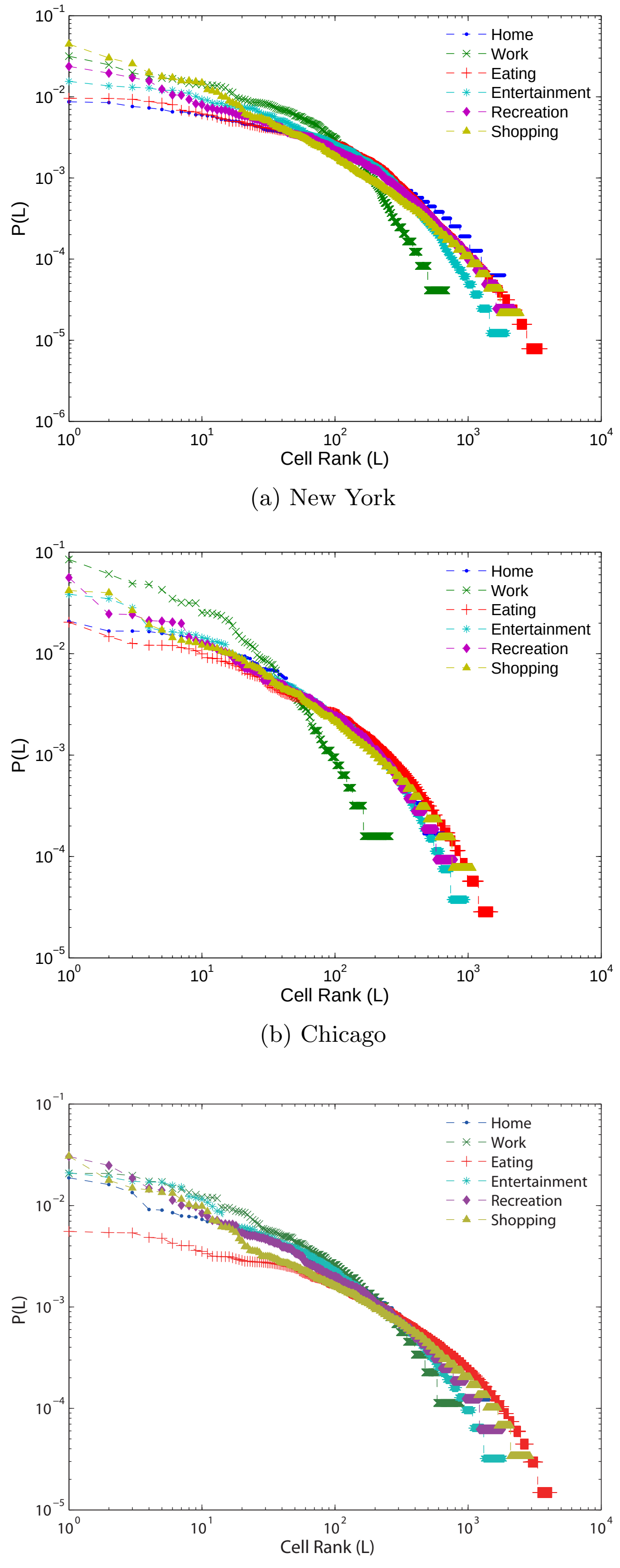
3. AGGREGATE SPATIAL ACTIVITY PATTERNS

3.1 Popular Places for Different Activities

To locate each check-in activity, a virtual grid reference is constructed by dividing the map into square cells of size $200 meters \times 200 meters$. We rank cells based on the number of check-ins for each activity category. For example, for a specific activity category, rank 1 represents the cell which has the highest number of check-ins for that activity category and so forth. We compute the frequency of check-ins for each of those ranked places. Figure 1 presents the frequency of visiting a place against its corresponding rank for each of the activity categories of the three cities: New York, Chicago and Los Angeles. The ranking pattern for different activity categories for different cities indicate that urban places are selected with diminished regularity. Furthermore, the regularity patterns follow a common scaling law as the distributions are fitted to truncated power laws $P(L) \sim L^{-\alpha} \exp^{-\lambda L}$. Table 3 presents the exponents fitting the truncated power law distributions for the three cities. The term $L^{-\alpha}$ dominates the distribution when the ranking L is small. Thus a larger α indicates a faster probability decrease when L is relatively small. Furthermore, each distribution has a cutoff value represented by $\frac{1}{\lambda}$ which captures the finite size of the activity locations for each activity category. Figure 1 shows how this cutoff value varies over activity categories for different cities. For instance, for New York City, these cutoff values are 116 and 588 for work and eating activities respectively indicating the number of cells where many people go for these specific activity purposes. Low values of the exponents indicate that below the cutoff values probability of selecting a cell does not vary significantly.

The ranking of a cell can be perceived as a measure of cell popularity, as higher ranking cells (smaller L) correspond to the places with higher number of check-ins indicating stronger ability to attract visitors. Preferential selection (a process where new objects tend to attach to popular objects) of activity locations exists in the ranking distribution for popular places. Popular places are more likely to attract both new and repeated visitors explaining the power law like curve (a straight line in log-log plot) of the distribution before the cutoff value. However, for cells with higher ranking (L) value, the probability decrease is much faster and resulting in a truncated power law distribution. There are several mechanisms that can explain the faster probability decrease for less popular cells. Schedule and distance constraints restrict the number of visits that a person can make or the less popular places may simply not be known by most people, so the preferential selection process fails in this case. Although this phenomena are very intuitive in the context of mobility behavior, our findings confirm that there is a remarkably simple scaling law explaining why few places in a city have most of the visitors.

Although the data for New York, Chicago and Los Angeles can be generally fitted into truncated power law distribution, however differences exist in the fitted exponents. It is observed that the λ parameters for Chicago are con-



(a) New York

(b) Chicago

(c) Los Angeles

Figure 1: Probability of a cell being visited by all the travelers against the corresponding rank of the cell

Table 3: Exponents fitting the truncated power law distributions

	New York		Chicago		Los Angeles	
Activity Category	α	λ	α	λ	α	λ
Home	0.3749	0.0020	0.4347	0.0051	0.4695	0.0021
Work	0.3797	0.0086	0.6899	0.0179	0.4897	0.0041
Eating	0.4722	0.0017	0.4732	0.0029	0.3853	0.0010
Entertainment	0.4558	0.0032	0.6029	0.0042	0.5914	0.0020
Recreation	0.5715	0.0018	0.6194	0.0040	0.6093	0.0016
Shopping	0.7781	0.0012	0.6709	0.0028	0.6356	0.0009

sistently larger than the other two cities. This is mainly due to fewer places that have check-ins and smaller size of dataset compared with the other two cities. Despite the effect of inflated λ parameters of Chicago, similarity exists for some activity categories across cities, such as recreation and shopping (large α and small λ), showing that the most popular cells attracts major proportion of visits. The differences of the exponents in the cell ranking distribution for different activity categories reveals how individuals choose to perform different activities in places with different popularity levels in the city, which indirectly reflect the unique urban characteristics.

3.2 Spatial Distributions of the Popular Places

We count the total number of check-ins for different activity purposes for each of this cells. The frequency for a cell for a specific activity purpose corresponds to the number of check-ins to that place. Figure 2 shows the check-in density for different activity categories of the Manhattan Island area in New York City. Figure 2a presents such a distribution where it indicates that people's home related visits are scattered over the city. As shown in Figure 2b, from the distribution of work related visits we find that there are not many check-ins at work related locations. However the work related visits are not as uniformly distributed as the home-related visits due to the concentrations of business locations at specific regions. In general it is found that home and work-related visits are scarce in social media check-in data compared to other human movement data (e.g. subway smart card transactions [16]).

Similar distributions for other activity categories can be derived by observing the frequency of visiting different places in the city for specific activity purposes(see Figures 2f-2e). These figures suggest that there are more check-ins for "other" (e.g. shopping, eating, entertainment and recreation) activity related visits than home and work-related visits. In general, there are few places that have a very high number of people that usually visit for shopping, eating, entertainment and recreation purposes; and the higher the frequency the more popular a place is. However these popular places are different depending on activity purpose. Furthermore, the distributions of check-in look different for various activity purposes. For instance the spatial check-in distributions for shopping (Figure 2f) and eating (Figure 2c) activities look very different. For shopping purpose check-in distributions are scattered all over the city while for eating purpose check-in distributions are concentrated within a specific area of the city.

3.3 Kernel Density Estimations of Spatial Distributions of the Popular Places

In section 3.2 we present the check-in density for different activity categories. In this section we adopt a non-parametric approach to estimate the check-in density distributions. To find the density of check-ins for each cell for a specific activity category, we use kernel density estimation technique [21] with a 2-dimensional Gaussian kernel and Silverman's rule for optimal bandwidth selection. Furthermore, to obtain time-dependent distributions, check-in data is split into different categories in 3-hour intervals, and probability density distributions are estimated for each case. Figure 5 in Appendix shows the kernel density estimation results for the Manhattan Island area in New York City. Four activity categories are presented, which are eating, entertainment, recreation and shopping, as these activity categories have apparent activity centers.

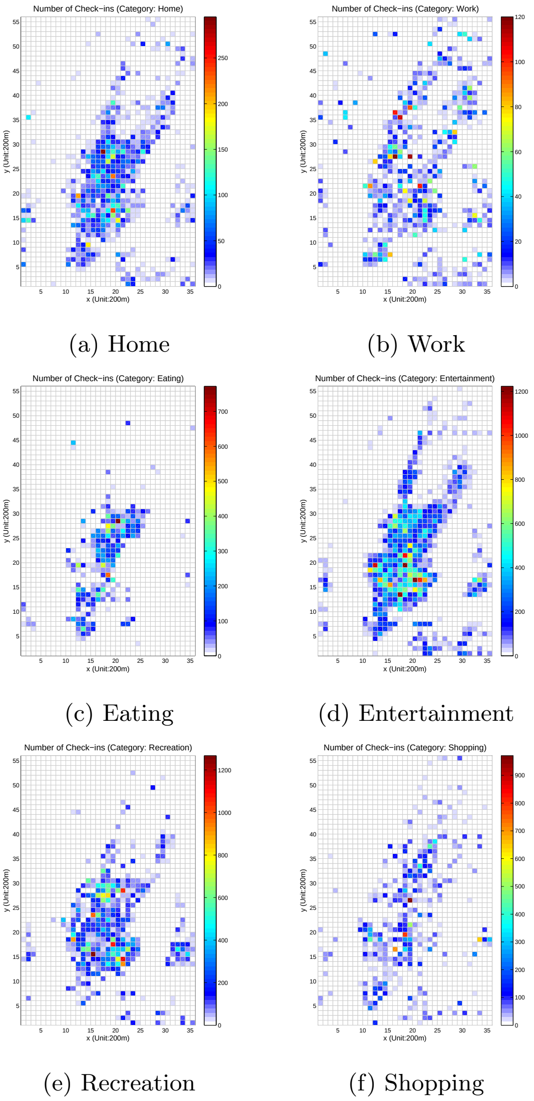
Compared with the grid maps (Figure 2), kernel density estimations provide more statistical information. The estimation results yield smooth distributions eliminating the local noise in certain degree. It also provides a non-parametric probability distribution integrating over all the sample space and with optimal bandwidth used to minimize the error between the estimated density and the true density. From the kernel density results, we can visualize the activity centers related to different activity categories at different time periods.

The kernel density estimation results reveal that the patterns for the evolution of activity centers can be classified into two groups: the first group is represented by eating and entertainment activity category, in which the activity center shifts from one region to another region as time elapses in a day; the other group is represented by recreation and shopping activity, in which activity centers seems remain stationary, since recreation sites like parks and shopping places like malls has consistent ability to attract visitors. The distinct patterns in urban activities are associated with the nature of different type of activities, and the information from kernel density estimation can help us to study the dynamic evolution of activity centers of each category in both space and time.

4. INDIVIDUAL MOBILITY PATTERNS

4.1 Temporal Mobility Patterns

To uncover the temporal regularity of urban human mobility, we investigate the distribution of visits for activity purposes at different hours of the day (see Figure 3). We also analyze the weekly rhythm of these visits. We observe that activity purpose has a pronounced impact on the time of activities. For instance eating activities has three distinc-



(a) Home

(b) Work

(c) Eating

(d) Entertainment

(e) Recreation

(f) Shopping

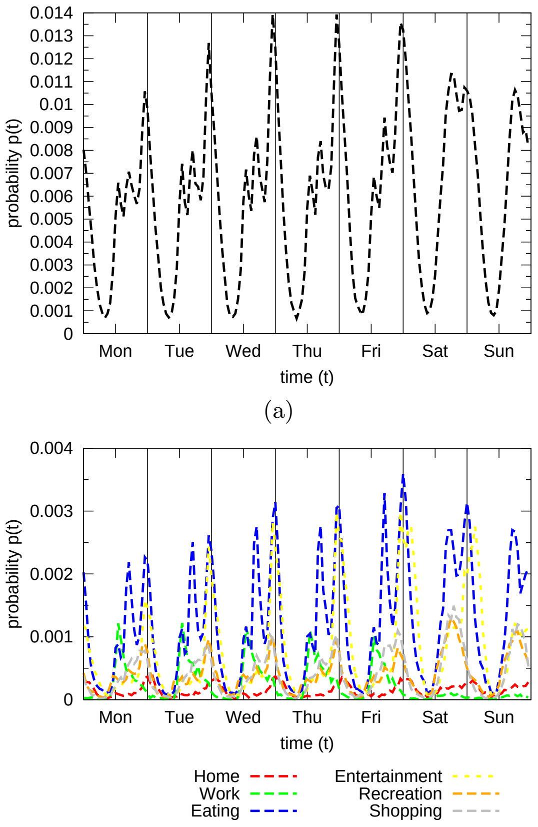
Figure 2: Check-in Density for Different Activity Categories

t peaks around noon (12pm), evening(6pm) and late night (11pm). Entertainment activities have peaks around late night as these visits mostly constitute of going to bars and night clubs. Weekly patterns suggest that shopping and recreation trips are predominant in the weekends.

4.2 Visitation Frequency

To find the probability of visiting a place we rank (L) each individual's visited places based on the number of times one visits the places over the study period. For instance, rank 1 represents the most visited place; rank 2 the second most visited place and so on. Then we calculate the frequency of each of these ranked places. Individuals are grouped based on the total number of different places they visit (N).

Figure 4 shows the probability of visiting different places against their corresponding ranks. People visit different places with diminished regularity. We observe that the distributions in Figure 4 follow a Zipf's law $P(L) \sim L^{-\eta}$ with an exponent that depends on the total number of visited locations. We find the coefficient of the Zip'f law $\eta \approx 1.2$ similar to the results from the mobile phone observations [20]. This resemblance is surprising given that there are certain kinds of activities (particularly work related visits) that are missing in the check-in data (see section 3.2).



(a)

(b)

Figure 3: Temporal Check-in Densities for a) all activities b) different activity categories

Previous study [16] found that most of the times people pay visits only to a few locations (two most visited places) and the probability of visiting the most visited place and the second most visited place are close to each other in value indicating most individuals' regular routine pattern of movements between their home and work location. However, since online social media users have less number of check-in activities at their homes and workplaces such pattern is not observed in our analysis.

5. CONCLUSIONS

This paper presents fundamental findings related to the spatio-temporal patterns of aggregate and individual mobility in a city using online social media data. Contrary to other mobility studies based on mobile phone call recordings, check-in observations and subway smart card transactions, we introduce activity category as a new dimension to our analysis. We first demonstrate how to characterize the temporal and spatial aspects of the mobility and activity patterns. From an aggregate perspective, it is found that people do not select their destinations randomly. Rather they select these places based on the popularity of the corresponding place; this means that, specific to an activity cat-

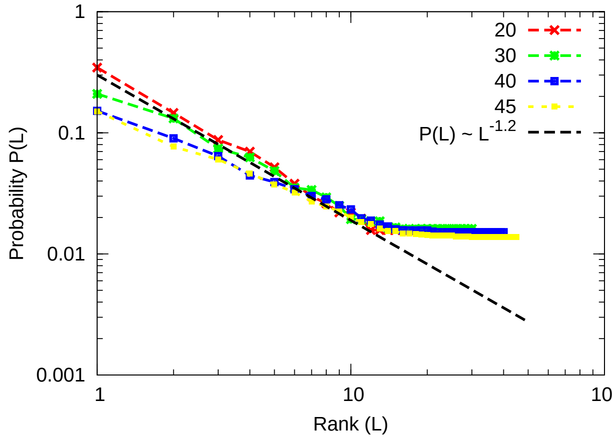


Figure 4: Probability of visiting different places against their corresponding ranking in log-log scale [P(L) vs. L]. The numbers in the legend refers to N, the total number of different locations visited by individuals. The straight line is a power-law decay with the exponent -1.2. The power-laws is shown as a guide to the eye.

egory, the more people select a place the more likely another person will select it. We discover a scaling law showing the relationship between the popularity of a place and the probability to select this place as a destination. We compare the relationship for different activity categories over three major cities in U.S. It is also found that the spatio-temporal distributions of check-in activities for different activity categories have distinct patterns. This implies a strong influence of urban contexts on peoples' activity participation and destination choices. Moreover, we observe different patterns for the evolution of urban activity centers in both space and time, which are closely related with the nature of the specific activity categories. In terms of individual-level patterns we observe that online social media users do not have many check-in activities in their homes and work places and users select places with diminishing probability following a Zipf's law. The exponent of the Zipf's law matches closely with the result from mobile phone studies. With the additional activity category information introduced in the analysis, the empirical findings from this study provide us richer insights on urban human mobility patterns.

6. ACKNOWLEDGMENTS

We are grateful to Professor James Caverlee from Texas A&M University for sharing the original Twitter check-in data. This work is funded by NSF Award 1131503 “Collaborative Research: From Warnings to Evacuation in Hurricanes: a Holistic Investigation using an Interdisciplinary Approach”.

7. REFERENCES

[1] Jie Bao, Yu Zheng, and Mohamed F. Mokbel. Location-based and preference-aware recommendation using sparse geo-social network data. In *ACM SIGSPATIAL GIS'12. Redondo Beach, CA, USA*, pages 199–208, November 2012.

[2] Zheng Vincent W., Yu Zheng, Xing Xie, and Yang Qiang. Collaborative location and acitivty recommendations with gps history data. In *WWW 2010, Releigh, North Carolina, USA*, pages 1029–1038, April 2010.

[3] Diego Saez-Trumper, Daniele Quercia, and Jon Crowcroft. Ads and the city: Considering geographic distance goes a long way. In *RecSys'12, Dublin, Ireland*, pages 187–194, September 2012.

[4] Zheng Yu. Location-based social networks: Users. In *Computing with Spatial Trajectories*, pages 243–276, 2011.

[5] Ling-Yin Wei, Yu Zheng, and Wen-Chih Peng. Constructing popular routes from uncertain trajectories. In *KDD'12, Beijing, China*, pages 195–203, August 2012.

[6] Reid Ewing and Robert Cervero. Travel and the built environment: A synthesis. *Transportation Research Record: Journal of the Transportation Research Board*, 1780:87–113, 2001.

[7] Kees Maat, Bert van Wee, and Dominic Stead. Land use and travel behaviour: expected effects from the perspective of utility theory and activity-based theories. *Environment and Planning B: Planning and Design*, 32(1):33–46, 2005.

[8] Bertil Vilhelmson. Daily mobility and the use of time for different activities . the case of sweden. *GeoJournal*, 48(3):177–185, 1999.

[9] Susan Hanson and O. James Huff. Systematic variability in repetitious travel. *Transportation*, 15(1):111–135, 1998.

[10] Andreas Schafer. Regularities in travel demand: An international perspective. *Journal of Transportation and Statistics*, 3(3):1–31, 2000.

[11] K W Axhausen, A Zimmermann, S Schonfelder, G Rindsfuser, and T Haupt. Observing the rhythms of daily life: A six-week travel diary. *Transportation*, 29(2):13–36, 2003.

[12] R Schlich and K W Axhausen. Habitual travel behaviour : Evidence from a six-week travel diary. *Transportation*, 30(1):13–36, 2003.

[13] Dick Ettema and Tanja van der Lippe. Weekly rhythms in task and time allocation of households. *Transportation*, 36(2):113–129, 2009.

[14] M. C. Gonzalez, Cesar A. Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[15] Brockmann, L Hufnagel, and T Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, Jan 2006.

[16] S. Hasan, C. M. Schneider, S. V. Ukkusuri, and M. C. González. Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics*, 151(1-2):304–318, 2012.

[17] Neal Lathia, Daniele Quercia, and Jon Crowcroft. The hidden image of the city: Sensing community well-being from urban mobility. In *10th International Conference, Pervasive 2012, Newcastle, UK*, pages 91–98, June 2012.

[18] Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Sui. Exploring millions of footprints in location sharing services. In *Proceeding of the 5th International AAAI Conference on Weblogs and Social Media*

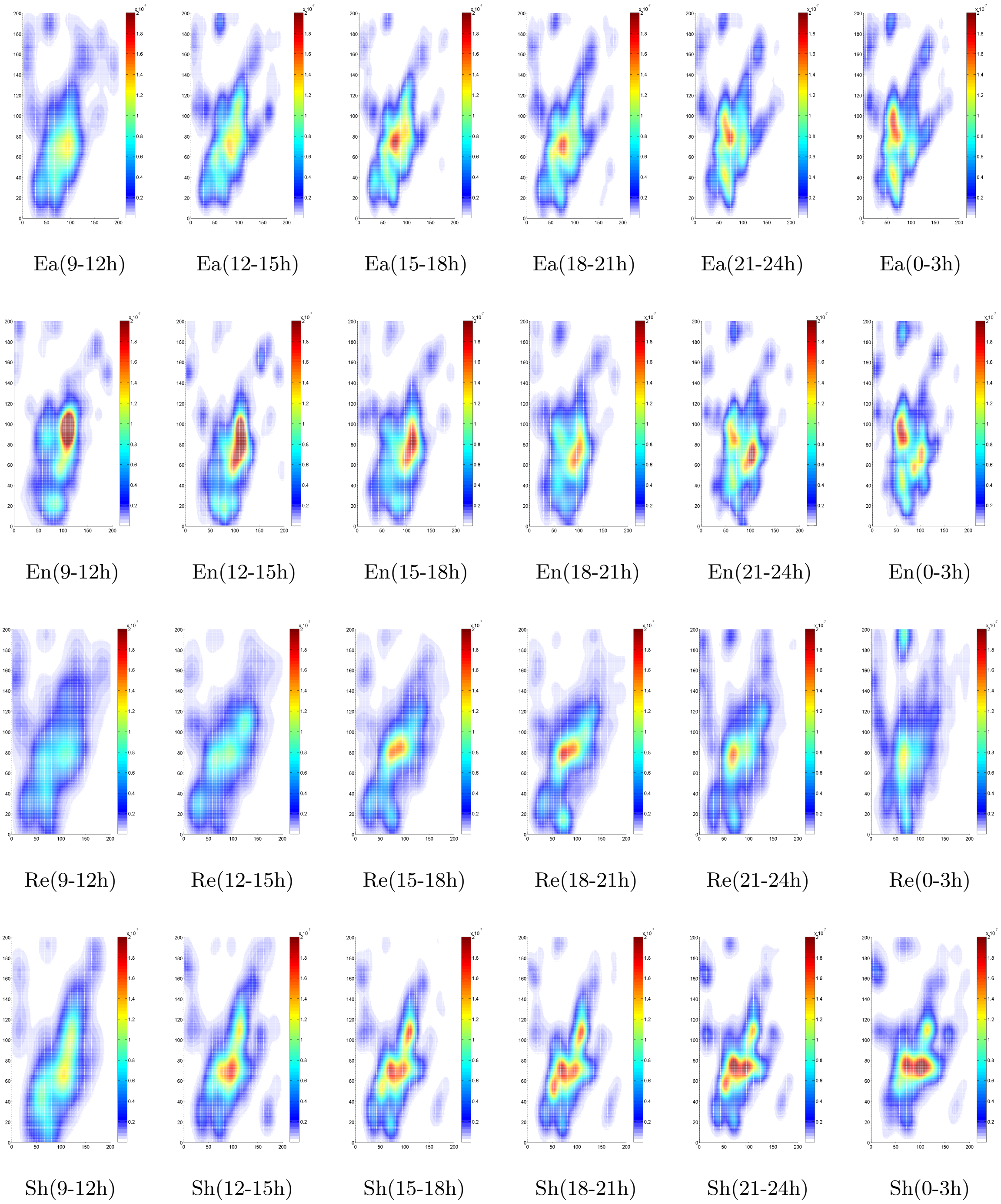
(ICWSM), July 2011.

[19] Eunjoon Cho, Seth A. Myers, and Jure Leskovec. Friendship and mobility: User movement in location-based social networks. In *KDD'11, San Diego, California, USA*, pages 243–276, August 2011.

[20] Chaoming Song, Tal Koren, Pu Wang, and Albert-Laszlo Barabasi. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, September 2010.

[21] Alex Ihler. Kernel density estimation toolbox for MATLAB. http://ssg.mit.edu/~ihler/code/, 2005.

APPENDIX



Ea(9-12h) Ea(12-15h) Ea(15-18h) Ea(18-21h) Ea(21-24h) Ea(0-3h)

En(9-12h) En(12-15h) En(15-18h) En(18-21h) En(21-24h) En(0-3h)

Re(9-12h) Re(12-15h) Re(15-18h) Re(18-21h) Re(21-24h) Re(0-3h)

Sh(9-12h) Sh(12-15h) Sh(15-18h) Sh(18-21h) Sh(21-24h) Sh(0-3h)

Figure 5: Kernel Estimation of Check-in Densities for Different Activity Categories. Ea- Eating En-Entertainment Re-recreation and Sh-Shopping activity. The numbers in the parentheses represent the start and end hour of the interval.