Not All Trips are Equal: Analyzing Foursquare Check-ins of Trips and City Visitors

Wen-Haw Chong, Bing Tian Dai, and Ee-Peng Lim
Dept. of Information Systems, Singapore Management University
80 Stamford Road, Singapore 178902
whchong.2013@phdis.smu.edu.sg, btdai@smu.edu.sg, eplim@smu.edu.sg

ABSTRACT

Location-Based Social Networks (LBSN) such as Foursquare allow users to indicate venue visits via check-ins. This results in much fine grained context-rich data, useful for studying user mobility. In this work, we use check-ins to characterize trips and visitors to two cities, where visitors are defined as having their home cities elsewhere. First, we divide trips into two duration types: long and short. We then show that trip types differ in check-in distributions over venue categories, time slots, as well as check-in intensity. Based on the trip types, we then divide visitors into long-term and short-term visitors. We compare visitor types in terms of popularities of check-in venues and proximities to friends' check-ins. Our findings indicate that short-term visitors are more biased towards popular venues. As for proximity to friends' check-ins, the effect is more consistently observed for long-term visitors. These findings also illustrate that locations of incoming visitors can effectively be analyzed using LBSN data in addition to conducting user surveys which are relatively costlier.

Lastly, we investigate the importance of visitor type information in models for venue prediction. We apply models including a state of the art kernel density estimation technique and ranking based on venue popularity. For each model, we consider two settings where visitor type information is absent/present. For long-term visitors, we observed little differences in accuracies. However, for short-term visitors, predictions are significantly more accurate by using type information. These findings suggest that venue prediction or recommender systems should consider visitor type to improve accuracy.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.2.8 [**Database Applications**]: Data Mining

COSN '15 November 02-03, 2015, Palo Alto, CA, USA

DOI: http://dx.doi.org/10.1145/2817946.2817958.

Keywords

check-in; visitors; Foursquare; long-term; short-term

1. INTRODUCTION

In recent years, Location-Based social Social Networks (LBSN) such as Foursquare and Yelp have grown rapidly in popularity. In particular, users can *check-in* with their mobile devices to various venues, thus providing researchers with a wealth of fine-grained data about visitation behavior.

In our current work, we are interested in the check-in behavior of travelers visiting a city away from their indicated home city. We simply term them as *visitors*. Visitors differ in their purpose, e.g. not all visitors are tourists or behave similarly. For better planning of city resources and promotion of tourism, some host countries/cities had conducted surveys on incoming visitors to determine their travel patterns and needs. One example is the survey on inbound visitors conducted by the Office of Travel and Tourism Industries (OTTI) of US Department of Commerce [1]. Such surveys are costly and do not always capture more fine grained mobility patterns of the visitors. In addition, they do not consider the social dimension of visits.

On the other hand, in the context of venue prediction for LBSNs, there has been little work that specifically studies the behavior of visitors considering their types and type-specific visit patterns [7, 14, 20, 9, 12, 18]. In particular, we note that for visitors who broadcast their check-ins while traveling, Foursquare provide sufficient data to estimate conservatively the trip duration or the lower bound. With the observed trip durations, one can now categorize trips into long or short and consequently, categorize visitors into long/short-term types. This leads to several research questions which we study via detailed empirical analysis on the Foursquare data collected for two major Asian Cities, Singapore and Jakarta. The **research questions** and our corresponding **contributions** are listed as follows:

1. *What are the differences between long and short trips?*
   (a) We showed that there are significant differences between long and short trips in their distributions over check-in venue categories.
   (b) Short trips have higher check-in intensity than long trips.
2. *What are the differences between long-term and short-term visitors?*
   (a) Short-term visitors tend to check-in at more popular venues, compared to long-term visitors.

(b) Check-ins for long-term visitors are slightly nearer those of his friends, compared to non-friends. For short-term visitors, this was only observed for one city.

3. *Does knowing the visitor type help to improve the accuracy of venue prediction?*

   (a) We can improve accuracy for short-term visitors if a prediction model is aware of their type.

   (b) For long-term visitors, we do not observe gain in prediction accuracy.

The outline of the paper follows the sequence of analysis steps that we have taken:

1. *Datasets construction*: We extract Foursquare check-ins and user profile data for the two cities being studied. See Section 2.1
2. *Trip categorization*: In Section 2.2, we extract trips from check-ins and categorize each trip as long/short.
3. *Visitor categorization*: Section 2.3 categorizes each visitor to a city as long/short-term based on his trip duration.
4. *Trip analysis*: Section 3 conducts empirical analysis on trips to contrast the differences between long and short trips.
5. *Visitor analysis*: Section 4 discusses our empirical analysis on visitors to contrast the differences between long-term and short-term visitors.
6. *Prediction experiment*: Lastly in Section 5, we apply models to predict check-in venues for visitors. For each model, we consider settings which include/exclude visitor type information. This ascertains the impact of visitor type on prediction tasks.

2. DATA AND CATEGORIZATION

2.1 Datasets

We study Foursquare check-in data collected for users visiting/residing in two Asian cities: Singapore and Jakarta, where Foursquare users are known to be highly active. We recognize that more cities can be studied to enhance this study. At the moment, we are limited to the two cities which we are collecting data on.

Besides check-ins, we also collect user profile information in Foursquare. Each user profile includes the *user-indicated* home city which we used to differentiate visitors from locals, as well as a list of friends. The friendship information is subsequently used when we analyze visitors' proximity to friends in terms of their check-in venues.

For Jakarta, we define a polygon based on the city boundaries and exclude the suburbs. This allows us to collect public check-in data that fall within Jakarta city. This step is not required for the island-state of Singapore, which is surrounded by the sea with limited entry points. Also note that Singapore is both a city and a country, hence visitors to Singapore are necessarily foreigners. In contrast, visitors to Jakarta comprise both foreigners and domestic travelers.

We apply a widely used method [8, 14] to collect check-in data, i.e. via crawling Twitter. As check-ins are publicly available only if the user broadcast them via Twitter, our data sets are gathered from related tweets. Each user is tracked throughout the study period, hence any of his check-ins outside the cities of interest are collected as well. This is necessary for us to estimate trip duration. In addition, for greater reliability in analysis, we only consider active users, which we define as having at least 10 public check-ins over the study period. Note that this is different from requiring a visitor to have at least 10 check-ins at his visited city.

For Singapore, the study period spans June 2013 to Nov 2014, comprising of 1,769,000+ check-ins, prior to filtering for active users. For Jakarta, 100,000+ check-ins are collected over a period of July 2014 to Feb 2015. Further statistics are listed in Sections 2.2 and 2.3 which discuss how we categorize trips and visitors.

2.2 Trip Categorization

To analyze the differences between long and short trips, we first need to extract and then categorize trips. The first step is straightforward. By tracking a user's check-ins over time, one can extract segment(s) where he check-ins at some given city of interest, say A, i.e. hence indicating trip(s) to A. Also recap that we define a user as a visitor to city A only if his listed home location is not in A.

For the second step, we need to categorize trips as long/short based on the trip duration. The simplest estimate is to use the time difference between the first and last check-in for a segment in A. However this requires the first and last check-in at A to be extremal: user check-ins at the moment of arrival and just before he departs from A. Otherwise the trip duration is underestimated. Obviously, it is also tricky to determine whether each user fits such a scenario.

To circumvent the described issues, we adopt a more general approach to estimate trip durations. Given two consecutive check-ins in two different cities, the crossing time is the time where the user crosses from one city to the other. This can be estimated as the mid-point of the two check-in times. It can be seen that a trip to a host city is necessarily bounded by two crossing times, the first being the arrival time at the host city and the second being the departure time from the host city. With the estimated arrival and departure time, the trip duration can then be estimated as the difference. This is a conservative estimate, not biased towards overestimating/underestimating the trip duration. Furthermore, it can be applied even if the trip contains only one check-in.

Formally, let the tuple $< t_j, C_j >$ represent a user's j-th check-in, occurring at time t_j at city C_j. As an example, assume the following sequence of check-ins involving cities A and B: $\{< t_j, B >, < t_{j+1}, A >, ..., < t_{j+m}, A >, < t_{j+m+1}, B >\}$. The trip duration for A is estimated as

$$(t_{j+m} + t_{j+m+1})/2 - (t_j + t_{j+1})/2 \quad (1)$$

Figure 1 presents a conceptual illustration.

Note that trip durations can be estimated only for uncensored trips, i.e. bounded by two crossings corresponding to the estimated arrival and departure times at the host city. After estimating the durations for all such trips, we apply two-mode Gaussian Mixture Modeling (GMM) to cluster trips into long and short trips. We also use GMM to derive a duration threshold[1], whereby out of sample trips with

[1]The threshold is an equiprobable point between the two modes using standard GMM formulation.

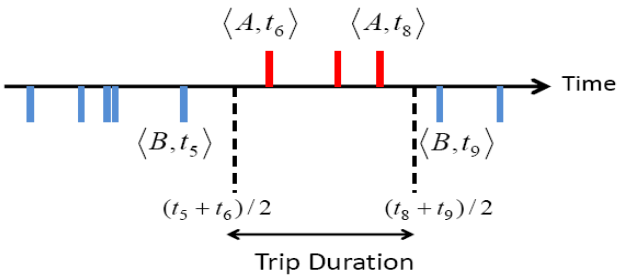


Figure 1: Estimating duration for a trip to city A. Red/blue tick marks are check-ins at city A/B.

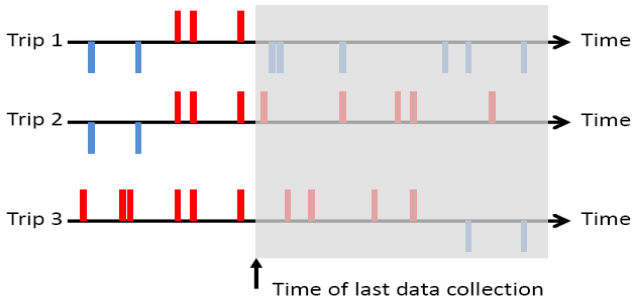


Figure 2: Red tick marks are check-ins during a trip. The shaded block is unobserved data. Trip 1 is short while trip 2 is long, but they are undifferentiable by their observed trip durations. Trip 3 is unambiguously a long trip as the observed trip duration is already long enough.

Table 1: Statistics from trips by visitors. Thresholds and durations are in days. Last row list the check-in count from locals.

	Singapore	Jakarta
Threshold (days)	9.9	8.34
Long trip mean duration	45.4	22.81
Short trip mean duration	2.8	2.24
No. of long trips (check-ins involved)	1,490 (37,124)	1,768 (6,708)
No. of short trips (check-ins involved)	2,976 (9,114)	5,350 (11,704)
Check-ins from locals	918,786	43,808

longer/shorter duration than the threshold are categorized as long/short.

Table 1 displays statistics derived for both Singapore and Jakarta, including check-ins generated by locals, i.e. from users listing their home locations as Singapore/Jakarta. In Section 3, we shall analyze the trips and local check-ins summarized in Table 1.

Reassuringly, the mean duration for short trips to Singapore is only slightly lower than the official statistics for average length of stay (3.5 days) for tourists [4]. The latter statistic excludes visitors from neighboring Malaysia, where short cross-border trips will have the effect of lowering the mean duration. For Jakarta, the city limits do not constitute a border between countries, hence the equivalent visitor statistics are not captured.

Instead of just long and short trips, one can also differentiate trips into more fine-grained duration categories. This will bring out larger differences when comparing extremal categories, e.g. very long versus very short trips. In the current work, we have used only two categories for brevity and to simplify our analysis. As will be evident in subsequent sections, this is already adequate for us to observe significant differences between long and short trips.

2.3 Visitor Categorization

Given a city A, we consider users to be locals if they indicate their home city in Foursquare as A, otherwise they are considered visitors if they check-in at A at least once. For locals, we only include those with at least half of their check-ins in A, thus excluding cases where locals are mostly staying elsewhere. For the visitors, we categorize them as long-term or short-term based on the following criterion:

- Short-term: The visitor has only short trips to A.
- Long-term: The visitor has at least one long trip to A.

whereby trips are categorized as described in Section 2.2. Furthermore, short trips are required to be uncensored and long trips can be censored. This is because if a trip is censored, the estimated duration is only a lower bound. It is then possible for a short trip to become a long trip as more data is collected. However if a trip is already of long duration, more data will not change the fact that it is a long trip. Figure 2 illustrates the concept.

It is possible to refine the criterion above to account for other factors, e.g. thresholding the number of trips such that visitors with too many short trips are treated as a separate category. We defer this to future work. For ease of analysis and brevity here, we categorize visitors into just long/short-term, showing that significant differences already exist.

Table 2: Count of visitors and locals. Ambiguous visitors have censored trips which appear to be short and no long trips, hence it is uncertain if they are short-term or long-term.

	Singapore	Jakarta
Long-term visitors	2,282	1,337
Short-term visitors	835	948
Ambiguous visitors	782	316
Locals	8,597	1,466

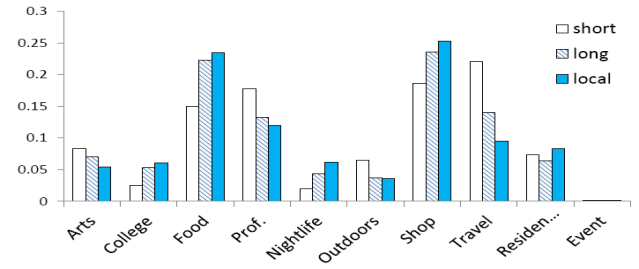
Table 2 displays the visitors/locals statistics for Singapore and Jakarta as gathered from Foursquare. We shall analyze these users in Section 4. We exclude visitors who are ambiguous, i.e. having censored trips which appear to be short and no long trips. With more data, their censored trips may well become long trips. Table 3 gives a breakdown of the number of check-ins per visitor type per city. For example, 388 short-term visitors to Singapore only made 1 check-in during their trip. As will be elaborated in Section 5.1, such sparse data impacts how we design personalized prediction models.

3. TRIP ANALYSIS

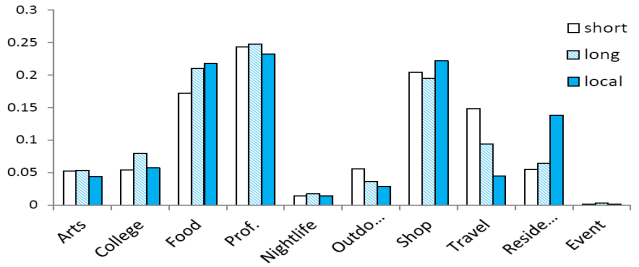
In this section, we study the differences between long and short trips by a number of measures including check-in distributions over venue categories/subcategories, check-in distributions over time and check-in intensity.

Table 3: Check-in distribution at the visited city per visitor type

	Singapore		Jakarta	
Check-in count	short-term	long-term	short-term	long-term
1	388	114	323	137
2	141	57	155	112
3	85	54	137	107
4	44	43	81	91
5	33	42	61	76
> 5	144	1972	191	814



(a) Singapore



(b) Jakarta

Figure 3: Check-in distributions over main categories for long trips (long), short trips (short) and from locals (local).

3.1 Venue Categories

Trips to a city are made for various purposes, affecting both the trip duration and the categories of venues visited. In this section, we examine the distribution over venue categories to understand trip purposes. Our task is facilitated by the fact that Foursquare already categorizes venues into 10 top-level categories which indicates their functions. These are: Arts & Entertainment (Arts), College & University (College), Food, Professional & Other Places (Prof.), Nightlife Spot (Nightlife), Outdoors & Recreation (Outdoors), Shop & Service (Shop), Travel & Transport (Travel), Residence and Events.

Figures 3(a) and 3(b) show the check-in distribution over venue categories for Singapore and Jakarta respectively. For example in Singapore, the probability of having a check-in from a long trip at a shopping venue (the 'Shop' category) is 0.24. The same probability is lower at 0.19 when the check-in comes from a short trip. For comparison, we also include the probability if the check-in is from a local.

From Figure 3, our key observation is: **long trips have check-in distributions more similar to that of local check-ins, when compared to short trips**. To quantify these differences, we compute the Jenson-Shannon divergence (JS) of distribution pairs as shown in Table 4. Note that the divergence values are small, but statistically significant, indicating that the differences are not due to chance. (We describe a test for significance in Appendix A.) In particular, divergence values between distributions from long trips and local check-ins are smaller for both Singapore and Jakarta, consistent with our key observation.

Table 4: Jenson-Shannon divergence values between pairs of category distributions. All values are statistically significant (p-value < 0.05)

JS divergence	Singapore	Jakarta
(long, local)	0.0047	0.0137
(short, local)	0.0363	0.0281
(long, short)	0.0185	0.0067

The key observation is intuitive since long-term visitors may have different focuses from short-term visitors or be assimilated to some extent in terms of check-in patterns. For example, if one stays for a long duration, shopping and dining needs tend to take on greater importance as compared to sightseeing or attraction hopping. There may also be a higher likelihood to visit places frequented by locals [27], instead of the usual tourist hangouts.

Besides the key observation, Singapore and Jakarta share other similarities:

- For 'Travel', short trips have much higher probabilities than long trips, which in turn have higher probabilities than local. This is intuitive since under this category, the various venue sub-categories are generally interesting to travelers, e.g. hotels, resorts, airports etc. (In Section 3.2, we shall examine the subcategories in more detail.)
- For both cities, long trips and local are also more similar in probabilities for 'Food'. This suggests that food places are more popular in long trips and among locals.

For observations specific to Singapore, long trips are closer in probabilities to local than short trips for 'Shop', 'Nightlife', 'College', 'Arts' and 'Prof'.

Lastly we note that the differences between long/short trips and local are more pronounced for Singapore than Jakarta. One contributing factor is the following: Singapore is both a city and a country, hence visitors are foreigners by definition. For Jakarta, visitors may be foreigners or fellow Indonesians residing elsewhere. We can expect the latter group to have somewhat similar visitation patterns/preferences to Indonesians residing in Jakarta. This brings down differences when we compare visitors and locals. For example, Indonesia is a Muslim majority country. One will expect most domestic visitors and locals to not visit nightlife venues where alcohol may be served. Consistent with this, Figure 3(b) shows little differences in 'Nightlife' probabilities between check-ins from trips and local. In contrast, Figure 3(a) shows that for Singapore, local check-ins have the highest probability for 'Nightlife', followed by long trips and with short trips last in place. As a side note, since visitors' ethnic composition affect their travel patterns, one can conduct interesting analysis of a city's ethnic composition or to quantify how cosmopolitan or mixed a population is. This may be useful as metrics for expat livability index for different cities [2].

3.2 Venue Subcategories

Earlier, we have seen that for check-in distributions over main venue categories, long trips are more similar to local check-ins, than short trips. As main categories are coarse and each can comprise many subcategories, we further analyze check-in distributions over subcategories as well. However it is not informative to display the complete distributions here due to the large number of subcategories (>700). Instead, we examine most probable travel-related subcategories where differences are more discernible.

Our procedure is as follows. First, for all trip types and local check-ins, we sort subcategories by probability. From the most probable 30 subcategories for short trips, we then manually select those that are travel-related and examine how their probabilities vary with trip types. The selected subcategories are either places of interest or provide transport and accommodation services required by typical travelers. Note that it suffices to select from short trips since we do not observe any travel-related subcategories that are probable for long trips / local but not among the top probable in short trips. We also observe the travel-related subcategories to be specific to cities, e.g. casinos have zero probabilities in our Jakarta data, but not in our Singapore data.

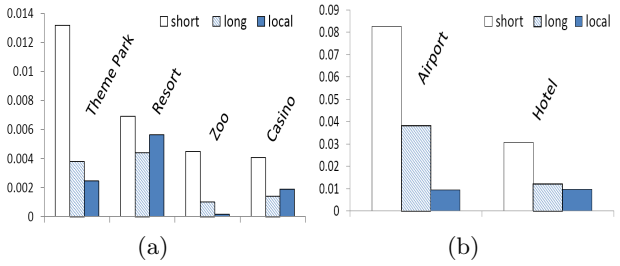


Figure 4: Travel-related subcategories (Singapore)

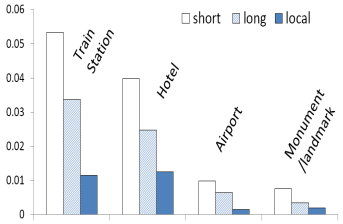


Figure 5: Travel-related subcategories (Jakarta)

Figure 4 displays the travel-related subcategories for Singapore. The subcategories are divided into the two figures 4(a) and 4(b) for better visibility due to differences in probability range. Figure 5 depicts similar information for Jakarta. From both figures, it is evident that short trips have higher probabilities in travel-related subcategories in both cities. Also consistent with the earlier observation for main categories, long trips have distributions that are closer to that of local check-ins, when we consider travel-related subcategories. Thus our key observation follows: **short trips have higher check-in probabilities for travel-related subcategories than long trips and local check-ins**.

For example short trips to Singapore has a check-in probability of 0.013 to theme parks[2], much higher than the probabilities of 0.004 for long trips and 0.002 for local check-ins. Interestingly, among the six subcategories displayed for Singapore, locals have the lowest probability for casinos. This is expected since the Singapore government imposes a levy of $100 [3] on local residents to enter casinos. The levy is exempted for foreigners.

From the most probable 30 subcategories for short trips, Jakarta has fewer travel-related subcategories (than Singapore). The inclusion of 'Train Station' warrants explanation since it was also probable for short trips in Singapore, but excluded. We include train stations as travel-related for Jakarta since visitors may arrive by train from other parts of Indonesia. This is rather different from train stations in Singapore, which we manually found to be referring to the local intra-city subway stations in many cases.

For Jakarta, Figure 5 shows the most probable subcategories i.e. 'Train Station', 'Hotel', 'Airport' and 'Monument/landmark'. For each subcategory, short trips have highest probability, followed by long trips and lastly, local. Thus short trips are again more biased towards travel-related subcategories.

Lastly, we observe that for both cities, locals do have some non-zero check-in probabilities at hotels, which is counter-intuitive. We attribute this to the trend of *staycations*, a form of in-country get-away, where one spends some nights at a local hotel for rest and relaxation. In fact, online searches of 'staycation' for both cities return a long list of hotels offering staycation packages to attract locals.

3.3 Check-in Probabilities over Time

We now compare the check-in probability distribution over hour of the day. Recap that we have observed long trips and local check-ins being more similar in category/subcategory distribution. For consistency, we now expect the mentioned pair to be more similar in temporal distribution as well (than short trips versus local check-ins).

Firstly, we compute the Jenson-Shannon divergence values between distributions in each city. This does not contradict what we expect, however the divergence values are not statistically significant. For example in Singapore, between short trips and local, we have divergence of 0.006 while between long trips and local, we obtain 0.0032. Nonetheless the divergence values are computed over all hours, which may obscure certain local differences. In the next paragraph, we describe differences that we observed by zooming in on certain hours.

Figure 6 presents the check-in probabilities over hour of the day for Singapore. It is clear that all distributions slightly peaked around lunch and dinner time. This is expected as we have earlier seen that the category 'Food' is very popular. On closer analysis, long trips and local check-ins appears more similar, especially around dinner time, i.e. 1800 to 2000 hours. This suggests a pick up in activities after office/school hours, leading to more check-ins. For short trips, one observes relatively more check-ins between lunch and dinner timings. Certainly, some of these check-ins would have been contributed by tourists who are not constrained by office hours, being free to spend the day visiting attractions, sight-seeing or shopping.

[2]Many check-ins are at the Universal Studios theme park, a popular attraction for tourists [5]

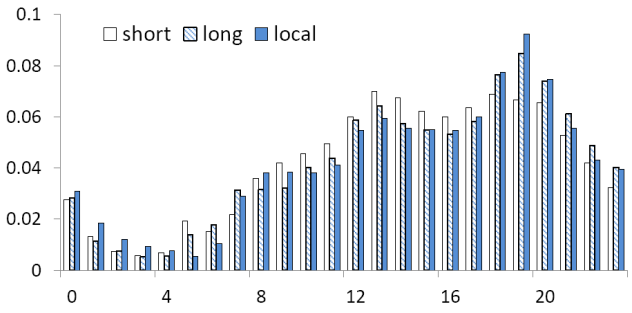


Figure 6: Check-in probability (Y-axis) over hour of day (X-axis) for Singapore.

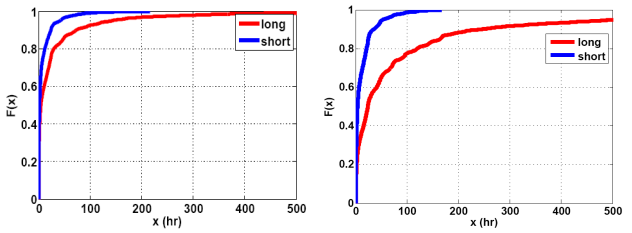
For brevity, we omit the plot for Jakarta, which is rather similar, except that check-in probabilities are lower than those of Singapore during the early hours (hours 0 to 5), probably due to less nightlife activities. For the check-in distributions over day of the week, the differences between trip types and local check-ins are small for both cities and we omit them from further discussion here.

3.4 Check-in Intensity

How frequent are check-ins made at city A when a visitor makes long/short trips to A? Our key observation is the following: **Short trips have higher check-in intensity than long trips at the destination, with smaller time gap between consecutive check-ins**.

Using trips with more than one check-in, we compute the time gap between consecutive check-ins and tabulate for different trip types. For short/long trips to Singapore, the time gaps are computed using 7,530/36,811 check-ins from 1,392/1,177 short/long trips. For short/long trips to Jakarta, time gaps are computed using 8,931/6,143 check-ins from 2,577/1,203 short/long trips.

Figure 7 plots the Cumulative Distribution Function (CDF) for time gaps. For both cities, it is clear that short trips have smaller time gaps, or equivalently higher check-in intensity. For example in Figure 7(a) for Singapore, around 93% of time gaps from long trips have duration less than 100 hours, whereas the corresponding figure for short trips is close to 100%. In Figure 7(b) for Jakarta, we have 78% of time gaps from long trips and 99% of time gaps from short trips to be under 100 hours.



(a) Singapore (b) Jakarta

Figure 7: CDF for time gap (hours) between consecutive trip check-ins. (Red:long trips, blue:short trips)

Time gap measurement requires more than one check-in in the trip segment, thus excluding a number of trips from analysis. However this does not affect our key observation. As a robustness check, we have computed another statistic: average time covered per check-in. For each trip, we simply divide the estimated trip duration by the check-in count, e.g. if a trip contains two check-ins over two days, each check-in covers one day on average. In this manner, all trips with at least one check-ins are also included. Results show that each check-in in short trips covers a shorter time duration on average, thus reaffirming our key observation. For Singapore, each check-in from short/long trips covers 1.7/11.1 days on average. For Jakarta, values for short/long trips are 1.6/12.5 days. The CDFs are similar in form as Figure 7 and omitted for brevity.

To understand the reasons behind the intensity differences, it is desired to conduct field studies or surveys of the visitors. We leave this to future work. Currently, we offer some intuitive reasons: short trips are more likely to be undertaken by tourists, who may visit more venues over a shorter period of time. Thus higher check-in intensity simply reflects more intense visitation activities. Another potential reason is that for short trips, one may tend to focus on key venues that are main draws since there is a need to maximize utility over limited time. Visiting such venues then increases one's propensity to check-in for the 'cool' factor or to enhance self-presentation [10]. Indeed, we shall see in Section 4.1 that short-term visitors (who make only short trips) tend to visit more popular venues.

4. VISITOR ANALYSIS

In this section, we conduct analysis on a visitor level. Since trip characteristics should carry over to visitors, we examine characteristics that are orthogonal to what we have studied previously in trips. For example, given that check-ins from short and long trips have different probabilities over categories, the resulting visitor types will naturally differ in this aspect as well.

Recap that we categorize visitors as long/short-term in Section 2.3. Based on this categorization, we now examine the differences between different types of visitors in terms of the popularity of check-in venues and their proximity to friends' check-ins. In summary, we observe:

- Short-term visitors are more biased towards popular venues than long-term visitors.
- For Singapore, both long and short-term visitors tend to check-in at venues closer to that of their friends than non-friends. For Jakarta, this proximity effect is only observed for long-term visitors.

4.1 Venue Popularities

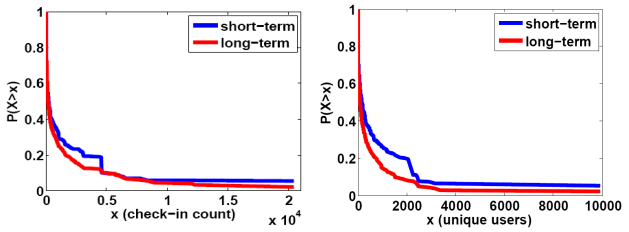
For each venue, we quantify its popularity using two measures: *check-in count* and *no. of unique visitors*. The two measures differ since each visitor can check-in multiple times at one venue. For each check-in instance from long/short-term visitors, we compute venue popularities. Figures 8 and 9 plot the Complementary Cumulative Distribution function (CCDF) for Singapore and Jakarta respectively.

Our key observation is the following: **short-term visitors tend to check-in at more popular venues than long-term visitors.** This is consistent across both cities and both popularity measures.

In both Figures 8 and 9, the CCDF curves for short-term visitors (blue) are above that for long-term visitors (red)

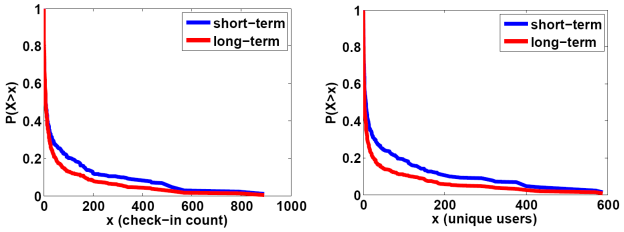
over a wide interval of popularity values (X-axis) thus supporting our key observation. For example, in Figure 8(b) for Singapore, 20% of check-ins by short-term visitors are to venues with at least 2000 unique visitors. For long-term visitors, the same proportion is only 8%. When we consider popularity in terms of check-in count, Figure 8(a) shows that 20% of check-ins by short-term visitors are to venues with at least 3000 check-ins. The corresponding proportion for long-term visitors is only 13%.

Venues are popular for various reasons, e.g. main attractions, must-try restaurants, key shopping areas etc. Given the limited time short-term visitors have, it is natural to focus on more popular venues where one is assured of a minimal level of utility or satisfaction. In contrast, long-term visitors will have more time and can afford to go off the well beaten track [27]. Such visitor behavior is well supported by the results here.



(a) Popularity (check-in) (b) Popularity (unique visitors)

Figure 8: CCDF for venue popularities for long-term (red) and short-term visitors (blue) to Singapore



(a) Popularity (check-in) (b) Popularity (unique visitors)

Figure 9: CCDF for venue popularities for long-term (red) and short-term visitors (blue) to Jakarta

4.2 Proximity to Friends' Check-ins

It has been established [7, 16, 12] that a user's friendships and his check-in venues are weakly related. Here, we investigate to what extent this is true for visitors. Our focus is different, since we specifically study visitors, who may not have deep social connections at the cities they are visiting. Our research question is: to what extent are check-ins and friendships related for short/long-term visitors.

Recap that we are using the friendship information that visitors have declared in their Foursquare profile (Section 2.1). We consider the case where visitor u to city A have friend(s) who also check-in at A. We do not differentiate between friends who are locals residing in A or other visitors. For u and each of his friend, we then compute the average distance between their sets of venues. This is repeated over all of u's friends, following which we take the mean to obtain a 'friend' statistic value for u. We compare this value with one from a 'non-friend' null model, where we replace u's friends with the same number of non-friends which are randomly sampled. Larger differences between the two values indicate that a visitor's check-ins and his friends' check-ins are more strongly related.

For a more rigorous approach that does not over-amplify the distance of non-friends, we restrict the null model sampling to nodes from the social network that u belongs to, based on visitor type. For example, if u is a short-term visitor, we sample from a social network comprising of all short-term visitors and their friends. Note that some users may be friends of both short and long-term visitors, thus existing in both short/long-term social networks. To collect more comprehensive 'non-friend' statistics, we sample for 10 trials such that each visitor has 10 non-friend values. Figures 10 and 11 compare the CDF for the 'friend'/'non-friend' values for Singapore and Jakarta respectively.

Figure 10 shows that for both visitor types in Singapore, visitors check-in at venues closer to that of their friends, when compared to randomly sampled non-friends. This is indicated by the CDF for 'friend' values (red) being consistently above that of 'non-friend' (blue). The effect of friends are rather similar for both visitor types in Singapore. Comparing the CDF median at $F(x) = 0.5$, the 'friend'/'non-friend' statistics are around 13/14 km in Figure 10(a) . For long-term visitors in Figure 10(a), at the median, 'friend'/'non-friend' statistics are around 9.6/10.5 km. Thus at the median, the distance reduction due to friends is around 1 km. It is also evident that the distance reduction is fairly constant for a wide band around the median.

Figure 11(a) shows that the distance reduction due to friends are barely discernible for short-term visitors to Jakarta. This contrasts with long-term visitors in Figure 11(b) where some reduction is observable. Thus, not all observations from Singapore carry over to Jakarta and city characteristics do affect whether visitors check-in close to that of their friends.

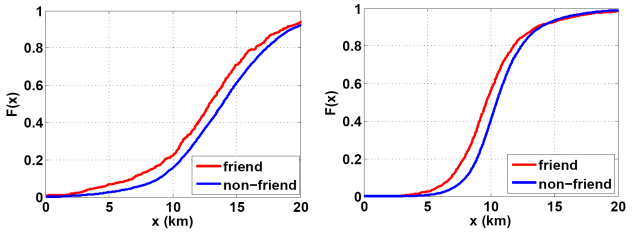
Following the analysis, we can now summarize our key observations as: **For long-term visitors, his check-in venues and friendships are weakly related. For short-term visitors, this was only observed for Singapore.**

Lastly, we point out that proximity to friends' check-ins is related to the homohily phenemenon which is driven by two processes: [15]: social influence and selection. Under social influence, visitors' check-in behavior are influenced by their friends. For selection, visitors with similar preference visit similar or nearby venues and bond with each other, thus becoming friends, e.g. nightclubbers connecting at clubbing sessions.

5. PREDICTION EXPERIMENTS

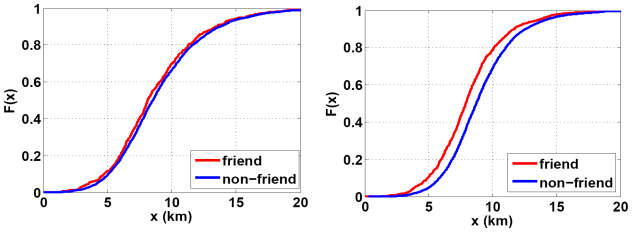
Based on each visitors' indicated home cities and trip duration, we have extracted and categorized visitors to Singapore and Jakarta into long/short-term. We have also shown that both visitor types have different characteristics. We now examine if the visitor type information can improve venue prediction accuracy for different models.

For each visitor, our task is to *predict the set of venues in the last 10% of his check-ins during the trip*. For example, if a visitor makes 20 check-ins during his trip to Singapore,



(a) Short-term visitors (b) Long-term visitors

Figure 10: CDF of average distance to venues of friends(red) /non-friends(blue) for Singapore



(a) Short-term visitors (b) Long-term visitors

Figure 11: CDF of average distance to venues of friends(red) /non-friends(blue) for Jakarta

we predict the venues involved in the last 2 check-ins. If a visitor has less than 10 check-ins for his trip, we simply predict for the last 1 check-in. Our results show that:

- **We can improve prediction accuracy for short-term visitors if a model is aware of their type**.
- For long-term visitors, it is not necessary to differentiate them from short-term visitors and we do not observe accuracy gains.

The reason for the second point depends on the prediction model employed, but for personalized models, the amount of personal data plays a part. As shown in Table 3, long-term visitors generally have more check-ins during their trips than short-term visitors. Hence if the model is able to exploit personal check-in history to make good predictions, then visitor type information is immaterial.

Inclusion/exclusion of visitor type information. To predict for each visitor, we use two experiment settings. In the first setting, the prediction models are aware of the visitor type and use only check-ins from the correct type for training or ranking. This can be seemed as a form of stratification in the hope of achieving better accuracies. In the second setting, the visitor type is unknown and the models simply utilized all visitor check-ins. For each setting, we use different models to rank venues per visitor such that high ranking venues are regarded as more likely to be check-in to. Henceforth we compare the results across settings per model and ascertain if the first setting gives better accuracy. Our prediction models include a sophisticated *Kernel-Density Estimation (KDE)* model adapted from a recent work [18] and simpler popularity-based ranking techniques.

In summary, for **Setting A**, our visitor-type aware prediction models are as follows:

- KDE model that includes a background component comprising check-ins from the specific visitor type
- Ranking of venues based on number of check-ins from the specific visitor type
- Ranking of venues based on number of unique visitors from the specific visitor type

In **Setting B**, the models are unaware of the visitor type:

- KDE model that includes a background component comprising check-ins from all visitors
- Ranking of venues based on number of check-ins from all visitors
- Ranking of venues based on overall number of unique visitors

Ranking venues based on check-ins/unique visitors are straightforward and self-explanatory. In the next section, we briefly explain the KDE model and the notion of the background component.

5.1 KDE Model

For each visitor in his city of visit, we fit a continuous two dimensional KDE model that estimates his check-in probability density at any city location. Compared to traditional spatial modeling techniques with Gaussian mixtures, the KDE model does not assume any parametric form for the spatial distribution and is better able to handle sharp transitions in spatial densities due to man-made or natural terrain.

We compute the probability of a visitor's check-in at a venue by predicting the density at the venue's spatial location: $e =< x, y >$. The density conditional on the training data $E = \{e_1, ..., e_n\}$, can be written as:

$$f(e|E) = \frac{1}{n}\sum_{i}^{n} \frac{exp[-\frac{1}{2}(e-e_i)^t C_{h(e_i)}^{-1}(e-e_i)]}{2\pi h(e_i)} \quad (2)$$

where we have used the Gaussian kernel with a diagonal covariance matrix, $C = \mathbf{I}h(e_i)$, and $h(e_i)$ is the local bandwidth for training point e_i, estimated by taking the distance to the k-th neighbor of e_i. Compared to using a global bandwidth, locally estimated bandwidths vary the degree of smoothing to better handle regions with sparse or high density of training points [18, 6].

5.1.1 KDE Models in Setting A

For a long-term visitor u, we can use solely his personal check-in history E_u for modeling. However to include additional information, we use a mixture of KDE components instead, (each component is equivalent to a KDE model on its own). Our choice of components differs from that of [18]. We include a component from the history of friends' check-ins $E_f(u)$, due to our analysis in Section 4.2. We also include a background component. In Setting A, the background component is estimated from the check-in history of other long-term visitors E_{L-term}. Formally, the KDE model contains 3 components:

$$\textbf{Long-term, Setting A: } f(e|E) = \alpha_u f(e|E_u) + \alpha_f f(e|E_{f(u)}) + (1-\alpha_u-\alpha_{f(u)}) f(e|E_{L-term}) \quad (3)$$

where 'α's are the mixture weights.

For short-term visitors, we again utilize a mixture of KDE components. However such visitors usually have very few check-ins at the visited city. For example Table 3 shows that

82.75%/79.85% of short-term visitors to Singapore/Jakarta have 5 or less check-ins at the visited city. This makes it difficult to build a KDE component from personal history. Hence, our KDE mixture uses only 2 components: friend's check-in history and a background component. As friends are specific to each visitor, there is still some personalization, although at much lower degree than long-term visitors. In Setting A, the background component is estimated from the history E_{S-term} of other short-term visitors. The mixture model is as follows:

$$\textbf{Short-term, Setting A: } f(e|E) = \alpha_f f(e|E_{f(u)}) + (1-\alpha_{f(u)}) f(e|E_{S-term}) \quad (4)$$

5.1.2 KDE Models in Setting B

The KDE models in Setting B differs from that of Setting A only in terms of how the background component is constructed. For both short and long-term visitors, the background component utilizes the check-ins of *all* visitors E_{All}, without any differentiation of visitor type. Other components in the respective mixture models are retained.

For long-term visitors, the KDE model is now:

$$\textbf{Long-term, Setting B: } f(e|E) = \alpha_u f(e|E_u) + \alpha_f f(e|E_{f(u)}) + (1-\alpha_u-\alpha_{f(u)}) f(e|E_{All}) \quad (5)$$

For short-term visitors, the model is written as:

$$\textbf{Short-term, Setting B: } f(e|E) = \alpha_f f(e|E_{f(u)}) + (1-\alpha_{f(u)}) f(e|E_{All}) \quad (6)$$

5.2 Experiment Design and Metrics

For each visitor, we hide check-in venues from the last 10% of his trips (by check-in count). We then build the KDE and popularity-based models to rank all candidate venues and assess ranking accuracies. Note that for model building, we only include friend and background check-ins that occur earlier than the last 10% of the trip. For long-term visitors, we only consider visitors with > 6 check-ins at the visited city such that the personal KDE component can be reasonably constructed. KDE mixture parameters, i.e. the 'α's are inferred using a grid search in step size of 0.05.

We refer to each visitor and his hidden venues as a test case. For Singapore, we obtained 1,498 test cases for long-term visitors and 582 test cases for short-term visitors. There are 65,701 candidate venues for ranking. For Jakarta, we have 461 test cases for long-term visitors and 382 test cases for short-term visitors. It is required to rank 30,254 venues in Jakarta. Note that due to the large number of venues per city, randomly ranking the venues will produce accuracies much lower than the KDE model or popularity ranking.

To measure ranking accuracy, each hidden venue is considered only once for the visitor even if he check-ins multiple times. We use the following accuracy metrics:

- **Mean Precision.** Given the p highest ranked venues for a visitor, precision at p, $Prec(p)$ is the proportion of hidden venues, i.e. venues that he actually check-in to. We then average precision over all test cases to obtain the mean precision at position p, $\mathbf{MP}(p)$.
- **Mean Recall.** Given the p highest ranked venues for a visitor, recall at p, $r(p)$ is the number of hidden venues retrieved at position p, divided by the total number of hidden venues. We then average recall over all test cases to obtain mean recall $\mathbf{MR}(p)$.
- **Mean Average Precision (MAP).** This is based on Average Precision (AP), commonly used in document retrieval tasks. For a test case, AP attains a perfect accuracy of 1 if all hidden venues are ranked higher than all other candidate venues. AP is computed as:

$$AP = \sum\nolimits_p Prec(p)\Delta r(p) \quad (7)$$

where $\Delta r(p)$ is the change in recall from position $p-1$ to p. We average AP over all test cases to obtain MAP. Also note that for each test case, we evaluate AP over all ranked venues (instead of just the top p).

5.3 Results

Tables 5 and 6 display the best prediction results in terms of MAP obtained for long-term and short-term visitors to Singapore. The corresponding KDE parameters are shown in Table 7.

5.3.1 Singapore

For each model, we compare the accuracies across settings A and B. For long-term visitors to Singapore, the relative differences are extremely small and negligible for both KDE and popularity ranking. In contrast for short-term visitors, all models perform much better in Setting A. For such visitors, the MAP gain of Setting A over B ranges around 100% for popularity ranking to 56% for KDE. Large gains are also observed for the other metrics of recall and precision. Hence for short-term visitors, it is beneficial to identify them as such, and use only check-ins from the correct visitor type. If one simply uses all visitor check-ins, there is too much noise from long-term visitors.

Evidently, check-ins from short-term visitors do not impose a problem of noise if included in the model for long-term visitors. There are several possible reasons, one of which is many check-in venues of short-term visitors are popular and also frequented by long-term visitors. For example, Singapore's main shopping belt is Orchard Road which attracts both short and long-term visitors (and locals). On the other hand, long-term visitors may frequent less accessible suburban malls, which draw fewer short-term visitors.

For personalized models such as KDE, the length of check-in history also plays a part. Short-term visitors spend much shorter duration at the visited city and many have insufficient check-ins for estimating the personal component (refer Table 3). On the other hand, long-term visitors have richer check-in history and the personal component plays an important role, i.e. weighted by α_u in Table 7. There is then less sensitivity to the background component, and in turn, to whether the background uses all visitor check-ins or not.

Table 7 displays the optimal KDE parameters. Interestingly for long-term visitors, the component for friends are not important, i.e. $\alpha_{f(u)} = 0$, which on the surface, seems contradictory to our earlier empirical analysis (Section 4.2). This can be explained by the fact that the visitor's personal check-in history has already captured the same information provided by his friends' check-ins. For example, if a visitor u and his friends frequent a shopping mall, u's check-ins alone may suffice for the KDE model to infer a high density value for that shopping mall.

5.3.2 Jakarta

Tables 8 and 9 display the results for long-term and short-term visitors to Jakarta, with corresponding KDE parame-

Table 5: Results for long-term test cases (Singapore)

Models					
Setting	MP(10)	MP(30)	MR(10)	MR(30)	MAP
KDE model					
A	0.0796	0.0346	0.172	0.201	0.1523
B	0.0798	0.0347	0.174	0.201	0.1524
Ranking venues by check-in count					
A	0.0496	0.0276	0.127	0.180	0.1118
B	0.0503	0.0277	0.132	0.181	0.1121
Ranking venues by unique visitor count					
A	0.0515	0.0282	0.134	0.197	0.1129
B	0.0482	0.0280	0.132	0.196	0.1130

Table 6: Results for short-term test cases (Singapore)

Models					
Setting	MP(10)	MP(30)	MR(10)	MR(30)	MAP
KDE model					
A	0.0320	0.0132	0.309	0.380	0.1748
B	0.0182	0.0072	0.174	0.206	0.1124
Ranking venues by check-in count					
A	0.0328	0.0143	0.316	0.411	0.1794
B	0.0155	0.0080	0.144	0.227	0.0823
Ranking venues by unique visitor count					
A	0.0332	0.0144	0.319	0.413	0.1836
B	0.0211	0.0087	0.200	0.247	0.0958

ters displayed in Table 10. Again, for each prediction model, we compare the results between settings A and B. Table 8 shows that for long-term test cases, Setting A does not consistently provide accuracy improvement across all models and metrics, when compared to Setting B. On the other hand, accuracies for short-term test cases (Table 9) are consistently higher for Setting A across all metrics and models. This agrees with our earlier observation for Singapore.

We note that for short-term test cases in Jakarta, Setting A provides a smaller magnitude of improvement over Setting B, as compared to Singapore. For example, MAP for KDE model increases by 10.39%, from 0.0635 to 0.0701, much less than the corresponding increase for Singapore (55.6%). One explanation is that both short and long-term visitors to Jakarta include fellow Indonesians residing outside the city. Such domestic visitors may have check-in behaviors that are more similar to each other, although trip durations may differ. Thus the differences between short and long-term visitors are reduced.

5.3.3 Comparisons

Comparing the optimal KDE parameters for Singapore (Table 7) and Jakarta (Table 10), the background component $1-\alpha_u-\alpha_{f(u)}$ for long-term Jakarta visitors has much smaller weights than the case for Singapore. Concurrently, long-term Jakarta visitors also have larger weights α_u for their personal history component than long-term Singapore visitors. Both observations suggest that long-term Jakarta visitors are relatively more personalized in their behavior.

To investigate further, we compute the normalized entropy of the distributions over venues for long-term visitor groups. We obtain 0.795 for long-term Jakarta visitors versus 0.74 for long-term Singapore visitors. Thus the former visitor group contains more uncertainty, which supports the notion of each visitor being more personalized. Various

Table 7: Optimal KDE parameters for visitors to Singapore

KDE model (long-term)			
Setting	α_u	$\alpha_{f(u)}$	$1-\alpha_u-\alpha_{f(u)}$
A	0.40	0.00	0.60
B	0.45	0.00	0.55
KDE model (short-term)			
Setting	-	$\alpha_{f(u)}$	$1-\alpha_{f(u)}$
A	-	0.15	0.85
B	-	0.20	0.80

factors may contribute to this, including city planning, car ownership and the availability of public transport etc. For example, the subway[3] is a popular transport mode in Singapore while Jakarta does not currently have a subway system. In the trivial extreme case, if all visitors only take the subway and check-in at venues near subway stations, then personalization is low.

Lastly, our results show that sophisticated models do not always outperform simpler techniques. While the KDE model easily outperforms ranking by popularity for long-term visitors (Tables 5 and 8), it fails to do so for short-term visitors (Tables 6 and 9). As mentioned earlier in the discussion of Singapore results and also shown in Table 3, such visitors have very few check-ins in their trip history for modeling, thus the KDE loses its advantage of being more personalized.

Table 8: Results for long-term test cases (Jakarta)

Models					
Setting	MP(10)	MP(30)	MR(10)	MR(30)	MAP
KDE model					
A	0.0245	0.0101	0.117	0.128	0.0928
B	0.0262	0.0106	0.131	0.138	0.0964
Ranking venues by check-in count					
A	0.0102	0.0062	0.0690	0.107	0.0352
B	0.0093	0.0066	0.0601	0.105	0.0318
Ranking venues by unique visitor count					
A	0.0102	0.0061	0.0671	0.109	0.0303
B	0.0111	0.0060	0.0724	0.109	0.0315

Table 9: Results for short-term test cases (Jakarta)

Models					
Setting	MP(10)	MP(30)	MR(10)	MR(30)	MAP
KDE model					
A	0.0128	0.0055	0.122	0.158	0.0701
B	0.0120	0.0041	0.114	0.116	0.0635
Ranking venues by check-in count					
A	0.0139	0.0072	0.132	0.203	0.0788
B	0.0126	0.0060	0.119	0.174	0.0703
Ranking venues by unique visitor count					
A	0.0136	0.0066	0.130	0.192	0.0769
B	0.0131	0.0067	0.124	0.195	0.0684

Remarks. We have shown that predictions are more accurate if we know a visitor is a short-term one, based on his trip duration, i.e. trip has ended. However, real applications require knowing the visitor type as early as possible, such

[3]www.lta.gov.sg/content/ltaweb/en/public-transport/mrt-and-lrt-trains.html

Table 10: Optimal KDE parameters for visitors to Jakarta

KDE model (long-term)			
Setting	α_u	$\alpha_{f(u)}$	$1-\alpha_u-\alpha_{f(u)}$
A	0.75	0.15	0.10
B	0.80	0.15	0.05
KDE model (short-term)			
Setting	-	$\alpha_{f(u)}$	$1-\alpha_{f(u)}$
A	-	0.05	0.95
B	-	0.00	1.00

that predictions can be made *during* the trip. For this, other data sources will be useful, e.g. content from tweets before or during the trip, immigration declarations etc. In other scenarios, we can motivate the visitors themselves to directly provide their trip duration to the prediction model. This is especially true for venue recommendation apps (related to venue prediction) in mobile phones, where the durations can be used to improve recommendations [25].

6. RELATED WORK

Users of location-based social networks have been well studied in prior work [13, 26, 7, 17, 16, 7, 14, 20, 9]. Gao and Liu [11] had provided a good survey in this area. However much focus has been on locally active users [16, 7, 14, 20, 9], which by excluding users with too few check-ins, may ignore most short-term visitors. In our case, we do not exclude short-term visitors even if they have only one check-in at their visited city. Some other works [17, 27, 19] studied inter-city visitation behavior, but do not differentiate between visitor types. In contrast, we study the differences between long/short trips and long/short-term visitors.

Liu et al. [19] define trips as city visits, which resembles our work. However trips are used differently to derive city-level interaction models as well as group cities into spatial communities. Wu et al. [24] define trips differently as each user's transition between different activity types, whereby activity types are derived from venue categories. They proposed a model for the transition probabilities of users.

Zhao et al. [27] matched interest communities (e.g. foodies) across regions, to recommend venues that are locally interesting for tourists, but omitted by dominant tourist resources. All city visitors are regarded as tourists, regardless of trip duration. As we have shown, short/long-term visitors have different check-in characteristics. Hence it will be interesting to compare recommendation accuracies across visitor types. We also envisage that numerous other recommendation/prediction models [7, 21, 17] are applicable and likely to perform differently on different visitor types. For example, topic models [14, 20, 13, 9] that model each user as a document and venues as words will encounter challenges for short-term visitors. Such users usually have little check-in history at the visited city and are analogous to extremely short documents.

Our analysis can be easily repeated on other forms of trajectory data, e.g. cell phone tower logs or GPS data, where trip durations can be estimated even more precisely. For GPS data, some related work includes [28, 29, 25], where the common goal is to make venue recommendations. In particular, Yoon et al [25] utilizes trip duration, input by users on their mobile devices, to make more useful recommendations. This supports our finding that visiting behavior is dependent on visitor type as determined by trip duration.

7. CONCLUSION

We have categorized trips and visitors to two cities and showed that significant differences exist between short and long trips and subsequently between short-term and long-term visitors. Our empirical analysis has been extensive and covers multiple aspects of check-in behavior. Many of the differences are intuitive and can be reasonably explained. For example, short-term visitors are biased towards more popular check-in venues as there may be a need to maximize utility over limited trip duration.

We follow up on the analysis by a venue prediction experiment. The results indicate that it is beneficial to identify short-term visitors properly and include that information in prediction models. Doing so increases prediction accuracy significantly. Equivalently, this indicates that trip duration and check-in behavior are highly related and that to make good predictions (or recommendations), one should factor in the trip duration. In certain scenarios, e.g. mobile apps, the app user himself can be easily motivated to input his intended trip duration to obtain better recommendations.

For further work, an interesting direction is to explore dynamic predictions as a trip progresses. In the context of a trip, the predictions may not only depend on the observed trip history to-date, but also on the estimated *remaining* trip duration. In fact estimating the latter is akin to the problem of *Survival Analysis*[23]. In survival analysis, one predicts the failure time of equipment or time of death of patients. In our problem domain, the failure time is analogous to the end of the trip. Check-ins and other data features are analogous to emitted symptoms which can help to refine the estimated trip end time. To our knowledge, it is still unexplored how one can estimate the trip end time and exploit this in a prediction model to achieve better prediction accuracies. At the moment, this seems to be a highly challenging problem, especially for short-term visitors. In fact the trade-off of latency versus accuracy in predictions arises [22].

8. ACKNOWLEDGEMENTS

This research is partially supported by DSO National Laboratories, Singapore; and the Singapore National Research Foundation under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority (MDA).

9. REFERENCES

[1] http://travel.trade.gov/research/programs/ifs/index.html.
[2] http://www.labourmobility.com/2011-top-cities-to-live-and-work-in-the-middle-east/.
html.
[3] www.ifaq.gov.sg/cra/apps/fcd_faqmain.aspx.
[4] www.singstat.gov.sg/publications/publications-and-papers/reference/mdscontent#tourism.
[5] www.yoursingapore.com/content/traveller/en/experience.html.
[6] L. Breiman, W. Meisel, and E. Purcell. Variable kernel estimates of multivariate densities. *Technometrics*, 19(2), 1977.

[7] C. Cheng, H. Yang, I. King, and M. R. Lyu. Fused matrix factorization with geographical and social influence in location-based social networks. *AAAI*, 2012.
[8] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui. Exploring millions of footprints in location sharing services. *ICWSM*, 2011.
[9] W.-H. Chong, B. T. Dai, and E.-P. Lim. Prediction of venues in Foursquare using flipped topic models. *ECIR*, 2015.
[10] H. Cramer, M. Rost, and L. E. Holmquist. Performing a check-in: emerging practices, norms and 'conflicts' in location-sharing using Foursquare. *MobileHCI*, 2011.
[11] H. Gao and H. Liu. Data analysis on location-based social netwoks. *Mobile Social Networking: An Innovative Approach, Springer*, 2014.
[12] H. Gao, J. Tang, and H. Liu. Exploring social-historical ties on location-based social networks. *ICWSM*, 2012.
[13] B. Hu and M. Ester. Spatial topic modeling in online social media for location recommendation. *Recsys*, 2013.
[14] K. Joseph, C. H. Tan, and K. M. Carley. Beyond âĂIJlocalâĂİ, âĂIJcategoriesâĂİ and âĂIJfriendsâĂİ: clustering Foursquare users with latent âĂIJtopicsâĂİ. *UbiComp*, 2012.
[15] D. B. Kandel. Homophily, selection, and socialization in adolescent friendships. *The American Journal of Sociology*, 84(2), 1978.
[16] A. S. H. Kautz and J. Bigham. Finding your friends and following them to where you are. *WSDM*, 2012.
[17] T. Kurashima, T. Iwata, T. Hoshide, N. Takaya, and K. Fujimura. Geo topic model: joint modeling of user's activity area and interests for location recommendation. *WSDM*, 2013.
[18] M. Lichman and P. Smyth. Modeling human location data with mixtures of kernel densities. *KDD*, 2014.
[19] Y. Liu, Z. Sui, C. Kang, and Y. Gao. Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. *PLoS ONE*, 9(1):e86026, 2014.
[20] X. Long, L. Jin, and J. Joshi. Exploring trajectory -driven local geographic topics in Foursquare. *UbiComp*, 2012.
[21] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. Mining user mobility features for next place prediction in location-based services. *ICDM*, 2012.
[22] R. Sen, Y. Lee, K. Jayarajah, A. Misra, and R. K. Balan. Grumon: fast and accurate group monitoring for heterogeneous urban spaces. *SenSys*, 2014.
[23] T. M. Therneau and P. M. Grambsch. Modeling survival data: extending the Cox model. *Springer-Verlag*, 2011.
[24] L. Wu, Y. Zhi, Z. Sui, and Y. Liu. Intra-urban human mobility and activity transition: Evidence from social media check-in data. *PLoS ONE*, 9(5):e97010, 2014.
[25] H. Yoon, Y. Zheng, X. Xie, and W. Woo. Smart itinerary recommendation based on user-generated gps trajectories. *UIC*, 2010.
[26] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. Magnenat-Thalmann. Who, where, when and what: discover spatio-temporal topics for Twitter users. *KDD*, 2013.
[27] Y.-L. Zhao, L. Nie, XiangyuWang, and T.-S. Chua. Personalized recommendations of locally interesting venues to tourists via cross-region community matching. *TIST*, 5(3):50:1-26, 2014.
[28] Y. Zheng and X. Xie. Learning travel recommendations from user-generated gps traces. *TIST*, 2(1):2, 2011.
[29] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from gps trajectories. *WWW*, 2009.

APPENDIX

A. SIGNIFICANCE TEST

We design a significance test for Jensen–Shannon (JS) divergence based on sampling. Given two probability distributions X and Y, JS divergence is defined as

$$JS(X||Y) = [KL(X||M) + KL(Y||M)]/2 \quad (8)$$

where $M = (X+Y)/2$ and KL(.) is the Kullback-âĂŞLeibler divergence.

If X and Y are not significantly different, both will be close to M. Thus we regard M as a form of null model and use this in our significance test. Each time, we draw 2 sets of samples from M and estimate the distribution per sample set. This gives a pair of distributions, from which we compute the JS divergence. Since both distributions are in fact generated from M, the divergence value can be interpreted as what is expected by chance given an identical distribution pair. We do this over multiple pairs and count the number of pairs with higher divergence values than $JS(X||Y)$. Such occurrences should be very low if X and Y are very different.

Formally, let X, Y be 2 empirical distributions estimated from samples of size S_x and S_y respectively, we test if they are significantly different via the following steps:

1. Compute $M = (X + Y)/2$, $d = JS(X||Y)$. Initialize counter $c := 0$.
2. For $i = 1$ to P
   (a) From M, draw 2 sets of samples, of sizes S_x and S_y.
   (b) For each sample set, estimate the multinomial distribution by proportions. Hence from 2 sample sets, obtain a pair of distributions: $M_{i,x}$, $M_{i,y}$.
   (c) Compute $d_i = JS(M_{i,x}||M_{i,y})$. if $d_i \geq d$, update counter $c := c + 1$.
3. Compute c/P. If this is less than α, then $JS(X||Y)$ is significant at p-value=α.

In all our significance tests, we have used $P = 1000$.