12-2015

# On neighborhood effects in location-based social networks

Thanh-Nam DOAN
*Singapore Management University*, tndoan.2012@phdis.smu.edu.sg

Freddy Chong-Tat CHUA
*HPLabs*

Ee-Peng LIM
*Singapore Management University*, eplim@smu.edu.sg

## Citation

DOAN, Thanh-Nam; CHUA, Freddy Chong-Tat; and LIM, Ee-Peng. On neighborhood effects in location-based social networks. (2015). *2015 IEEE / WIC / ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT): Proceedings, Singapore, December 6-9*. 477-484.
**Available at:** https://ink.library.smu.edu.sg/sis_research/3463

# On Neighborhood Effects in Location-based Social Networks

Thanh-Nam Doan
School of Information Systems
Singapore Management University
tndoan.2012@phdis.smu.edu.sg

Freddy Chong-Tat Chua
Mechanisms and Design Lab
HPLabs, Palo Alto
freddy.chua@hp.com

Ee-Peng Lim
School of Information Systems
Singapore Management University
eplim@smu.edu.sg

***Abstract*—In this paper, we analyze factors that determine the check-in decisions of users on venues using a location-based social network dataset. Based on a Foursquare dataset constructed from Singapore-based users, we devise a stringent criteria to identify the actual home locations of a subset of users. Using these users' check-ins, we aim to ascertain the neighborhood effect on the venues visited, compared with the activity level of users. We further formulate the check-in count prediction and check-in prediction tasks. A comprehensive set of features have been defined and they encompass information from users, venues, their neighbors, and friendship networks. We next propose regression and classification models to address the two prediction tasks respectively. Our experiments have shown that the two models especially the classification models outperform the baseline methods when all features are used. We also analyze feature importance and found that despite their similarity, the two prediction tasks actually require different weights on the features as learned by the regression and classification models. Finally, it was found that user's home location for deriving user-venue distance feature is a better feature than user's center of the mass.**

## I. Introduction

**Motivation.** With the wide use of smartphones and tablets, many users today are attracted to use location-based social networks (LBSN) to share information about their visits to different locations in their own cities or other parts of the world, to give comments on these visited locations, to search for locations, and to interact with friends. Each mention of visit to some location is often known as a *check-in*. By analyzing all the check-in data of a user or a group of users, one can then derive interesting insights about users' check-in behaviors.

There are several important applications that can benefit from these insights and they include location recommendation, user profiling, and business intelligence. The insights also allow city planners to design better public transportation systems and housing plan to meet residents' needs. Businesses can leverage on the insights to determine suitable store locations.

We believe that the home locations of users and locations of venues affect the way users perform their visits or check-ins. Intuition tells us that people living in a city are less likely to visit other cities compared with places within their home city. Such an intuition should also apply at the fine-grained neighborhood level. Neighborhood information comes in two forms: (a) user's neighborhood, and (b) venue's neighborhood. While every venue's location is known and static, the user's location is dynamic and the determination of the user's home location is itself a research problem. This explains why previous research on location-based social networks did not study users' check-ins considering all aspects of neighborhood effect.

**Research Objectives.** In this paper, we analyze a Foursquare dataset that consists of almost one year of check-in data by Singapore-based users. We focus on determining the neighborhood effect on these users' check-in behavior. To do that, we first need to determine the home locations of a subset of these users. We seek to answer the research questions of (a) how likely a user will perform check-ins on places nearby his home versus far away places, and (b) how likely two neighbors will share check-ins places.

We also propose to address two closely related but different check-in activity prediction tasks. The first task predicts the number of check-ins a user performs on a venue. The second task predicts if check-in on a venue will be performed by a user. In these tasks, we want to determine features that can allow us to produce accurate prediction results.

In the following, we summarize our results and findings:

- We carefully collected a set of publicly available check-in data which comprises the check-in behavior of users from Singapore and determine the home locations of a subset of them through some stringent criteria. This gives us a good user dataset to embark on this research.
- We conduct some analysis of the check-in behavior of the dataset to determine the neighborhood and user/venue popularity effect on check-ins.
- We propose a taxonomy of features covering user, venue, user/venue neighborhood information which are subsequently used in the check-in count prediction and check-in prediction tasks.
- In our experiments, we show that our proposed supervised methods generally perform better than unsupervised baseline methods. We also show that user's home location for deriving user-venue distance feature is a better feature than user's center of the mass. Users who are active in performing check-ins, venue popularity and venue distance from user's home location, alone or combined with other features, can significantly affect the prediction results.

**Paper Outline.** The rest of the paper is organized as follows. Section II summarizes related works. Section III shows the insight of our dataset and Section IV defines the check-in activity prediction tasks and our proposed feature set. Section V covers the experiment setup and evaluation of our proposed methods. The analysis of feature importance will be given in Section VI. Section VII will conclude the paper and offer some directions for future works.

## II. Related Works

Our work is related to two bodies of research works. The first body of works focuses on analyzing check-ins made by users in LBSN, e.g., Foursquare. The second body of works performs prediction of user movement.

**Check-in analysis in location-based social networks.** Anastasios *et al.*, using a large number of foursquare check-ins [18], found that users demonstrate different temporal patterns of check-ins at different types of venues on weekdays and weekends. 20% of the consecutive check-ins are found within the distance of 1 km, while much smaller proportion of such check-ins are found larger than the distance of 10 km.

Cramer *et al.* conducted a user-study to understand the check-in behavior of users in Foursquare and to determine the motivation behind location sharing among users [7]. Beyond check-in behavior, Vasconcelos *et al.* studied their behaviors of posting tips, dones and to-dos [22]. The work also clustered users into profile types (e.g., influential users, spammers, etc.) according to their behaviorial patterns.

**Home location identification.** Prior to the era of LBSN, researchers heavily depends on GPS data from phones to determine home locations [21]. With the logged user mobility data, Krumm attempted to predict the home locations of users by using heuristics rules [11]. Using the users' home locations as input for web search engine, it is shown that the user names may be compromised. In our context, the home locations of users are self-reported instead of using heuristics rules which may not be accurate.

**Check-in prediction research.** Much research has been done in mining user trajectories from GPS data [10], [12], [13], [24]. Such device-tracked movement data is quite different from self-reported movement data found in LBSN such as Foursquare[1] or Facebook Places [2].

Chen *et al.* [5] proposed the use of matrix factorization with multi-center Gaussian model to recommend users new venues. They used social information as the regularization. However, the work does not consider user and venue neighbors in the recommendation approach. Cho *et al.* [6] viewed check-in locations of users as the mixture of check-ins near *home* and *work*. They further proposed Bayesian models to predict the check-ins using time and social network features. Our differences are that we do not consider the time of check-ins and our method assumes one home location for each user.

Gao *et. al.*[9] addressed the cold start problem of predicting a user checking into a new venue under the effect of neighbors and friends on social network. In our research, we also consider both distance and social network factors in check-in prediction. We further consider neighborhood of previously visited venues as another factor in our prediction. Noulas *et al.* [17] is the first work which tried to predict the next move of users. However, this work and others [14], [1], [4], [23], [19] do not consider the impact of neighbors of users to their check-in behaviors.

## III. Check-in Data Analysis

### A. Dataset

We first crawled a Foursquare dataset (denoted by **FQ**) that consists of 1.11 millions check-ins by Singapore users who publish their check-ins on public Twitter stream between August 15, 2012 and June 3, 2013. As shown in Table I (second column), this dataset consists of 55,891 users and 75,346 venues in Singapore. These users declare Singapore to be their profile location.

TABLE I: Dataset Statistics

| | FQ | H_FQ |
|---|---|---|
| # users | 55,891 | 856 |
| # venues | 75,346 | 12,020 |
| # check-in's | 1.11M | 63,777 |
| # user-venue pairs with $>$ 0 check-ins | 541,588 | 28,298 |

From this dataset, we then identify a subset of users whose home locations can be determined. This subset of users and their check-in data form a smaller dataset denoted by **H_FQ**. We describe the home location identification method in the next subsection.

### B. Home Location Identification

Home location could influence a user's activity region. For example, people normally visit the supermarkets near home for grocery shopping, attend schools and patronize fitness facilities in the home neighborhoods. Home location could also infer the social status of a user (e.g., living in luxurious apartments versus public housing) that could be strongly correlated with the purchase patterns and thus check-in behaviors. While the **FQ** dataset covers all check-in data from a set of users, it has no information about the users' home locations.

In this research, we therefore select a subset of users whose home locations can be clearly identified using both their check-ins and check-in messages. The following are the detailed steps:

- We select a subset of venues under the "home (private)" category which is in turn a sub-category of the "residence" category. There are 8,447 venues satisfying this criteria. There are 74,944 check-ins on these venues by 5,199 users. At this point, it is still unclear if these venues are the home locations of 5,199 users.
- We further select a subset of 3,276 users who have checked in at only one "home (private)" venue. This rules out users who have multiple "home (private)" venues.
- We finally select an even smaller set of users who also shouted some home relevant messages during check-ins. We use a set of "home" related key phrases, e.g., "back home", "home finally", etc., to identify such messages. As long as any of the key phrases are found, the check-in location is used as the home location.

[1] https://foursquare.com/

[2] https://www.facebook.com/places/

We finally obtained a dataset which includes 856 users and their home locations. We call this dataset **H_FQ**. These users have 63,777 check-ins on 12,020 venues as shown in Table I. Note that this represents 1.5% of all users and 5.7% of all tweets in **FQ**. As a user can have multiple check-ins on the same venue, the number of unique user-venue pairs with non-zero check-ins is 28,298.

**Center of the mass.** For evaluation purposes, we also define the center of the mass of each user as only a small fraction of users have home locations. Suppose that user $u$ has performed check-ins at $n$ venues and we denote the location set of these venues as $S = \{(lat_i, lng_i) | i \in \mathbb{N} \wedge 1 \leq i \leq n\}$ where $(lat_i, lng_i)$ represents the latitude and longitude of each check-in venue $i$. The center of the mass of $u$ is defined by $(lat_u$,$lng_u)$ where $lat_u = \frac{\sum_i lat_i}{n}$ and $lng_u = \frac{\sum_i lng_i}{n}$.

Center of the mass is often not the same as true home location unless the user performs check-in at home location only, which is very rare. Center of the mass may also likely return a location where no venue can be found. In Section IV, we will derive features from both user's home location and center of the mass to ascertain their utility in predicting missing check-ins.

### C. Check-in Venue Category Analysis

Foursquare has a three-level category hierarchy for venues[3]. Each venue is assigned to one of the nine level-one categories, i.e., *food*, *shop & service*, *professional & other places*, *travel & transport*, *nightlife & spot*, *arts & entertainment*, *college & university*, *outdoors & recreation*, and *residence* as shown in Table II. In addition, zero or more detailed level-two or level-three categories under the already assigned level-one category may be further assigned to the venue. For example, a food venue can be assigned level-two category, say Indian restaurant, as well as a level-three category, say North or South Indian Food. In this work, we use the level-two category label to determine venue similarity. Two venues are considered as similar if they share any same level-two category.

The statistics of the **H_FQ** dataset is shown in Table II. We leave out the level-three categories since they are too fine-grained for this research. The largest level-one categories by number of level-two categories are *shop & service* and *food*. Nevertheless, they are not the largest by other measures. The *food* and *professional & other places* categories have most number of venues. The *food* category has most number of users.

### D. Neighborhood Effect on Check-in Activities

**Analysis of distance between users and visited venues.** It has been reported in previous works [6] that users are most likely to visit locations nearer to their home locations than far away locations. These works often involve the predicted home locations of users instead of user-reported home locations. In the following, we report the analysis using the distance between check-ins and actual home locations in **H_FQ** dataset. We also present the results at the user level.

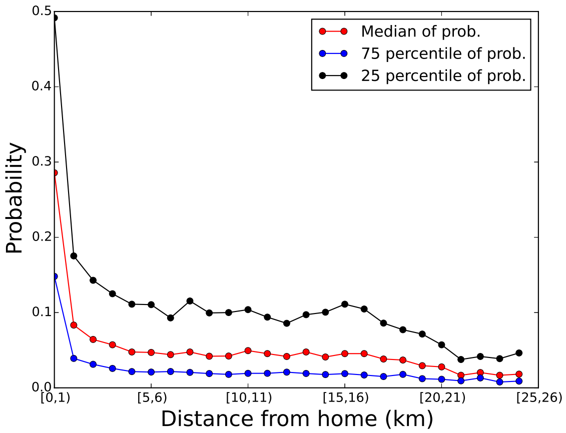


Fig. 1: Fraction of check-ins as a function of distance from home in **H_FQ** dataset.

For each user, we bin her check-ins according to the distance from the user's home location. Every 1-km distance is a bin and we compute the probability of check-ins within each bin by dividing the number check-ins within the bin by the total number of check-ins of this user. As shown in Figure 1, the median, 25-percentile and 75-percentile probabilities of check-ins decreases with the distance. The maximum distance from home location to venue is 36.7 km as this is almost the largest diameter of the city. As the large distance bins involve the check-ins of very few users, we remove the bins with distance larger than 25 km. Based on this result, we could conclude that users are more likely to perform check-ins near their home locations.

**Analysis of visited venues between neighbors.** When two users' home locations are near to each other, there could be similarity between their check-ins that can be attributed to the similar daily patterns shared by people living in the same neighborhood. Cho et. al earlier showed that periodicity and friendships can have effect on users' check-in locations [6]. The work however did not consider neighborhood effect.

We ascertain this neighborhood effect using the **H_FQ** dataset. Empirically, we define two users to be neighbors when the distance between their home locations is less than 100 meters. We found that the average Jaccard similarity between the venues visited by a user and his $x_i$ neighbors is 0.0105. Compared with 0.005, the average Jaccard similarity between a user and randomly selected $x_i$ users, the user clearly shows more similarity in the visited venues with his neighbors than with strangers. From this result, we could conclude that the effect of neighborhood is two times stronger than the random one. Hence, neighborhood effect should be included into the process of predicting check-ins.

Figure 2 shows the average Jaccard Similarity of check-in venues between pairs of users with different inter-home distance. We first calculated the inter-home distance and Jaccard Similarity of check-in venues of every pair of users in **H_FQ**. We then group pairs of users into distance bin of 1 km. For example, the first bin contains all user pairs whose distance is less than 1 km. The second bin contains user pairs whose distance is between 1 km and 2 km. We exclude those

[3] https://developer.foursquare.com/categorytree

TABLE II: Category Statistics of **H_FQ** Dataset

| Level-1 cateogry | Food | Shop & Service | Professional & Other Places | Travel & Transport | Nightlife & Spot | Arts & Entertainment | College & University | Outdoors Recreation | Residence |
|---|---|---|---|---|---|---|---|---|---|
| # sub-categories | 59<br>(19.67)% | 64<br>(21.33)% | 42<br>(14.00)% | 25<br>(8.33)% | 16<br>(5.33)% | 31<br>(10.33)% | 27<br>(9.00)% | 35<br>(11.67)% | 1<br>(0.33)% |
| # venues | 3,657<br>(31.98)% | 1,654<br>(14.47)% | 2,302<br>(20.13)% | 1,085<br>(9.49)% | 380<br>(3.32)% | 438<br>(3.83)% | 553<br>(4.84)% | 509<br>(4.45)% | 856<br>(7.49)% |
| # users | 689<br>(13.05)% | 696<br>(13.18)% | 856<br>(16.21)% | 624<br>(11.82)% | 279<br>(5.28)% | 493<br>(9.34)% | 363<br>(6.87)% | 425<br>(8.05)% | 856<br>(16.21)% |
| # check-ins | 11,501<br>(18.96)% | 11,501<br>(18.96)% | 10,337<br>(17.04)% | 8,781<br>(14.48)% | 1,840<br>(3.03)% | 2,577<br>(4.25)% | 4,218<br>(6.96)% | 1,645<br>(2.71)% | 8,247<br>(13.60)% |

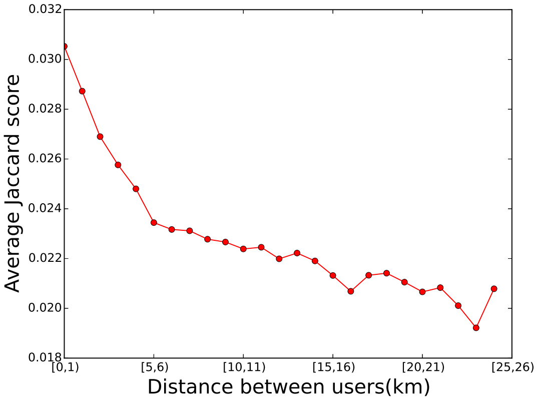


Fig. 2: Relationship between Jaccard score and distance between every users in **H_FQ**.

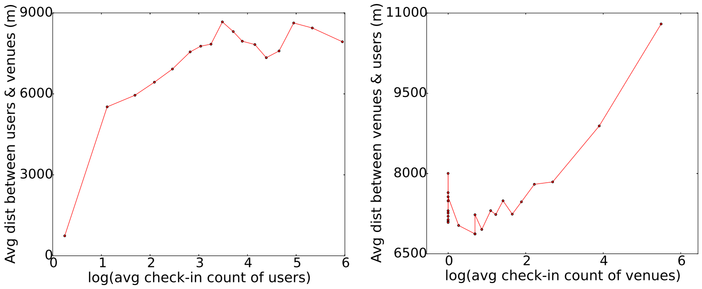
user pairs with distance larger than 25 km as they are few in number. Figure 2 shows that the average Jaccard Similarity decreases with inter-home distance. Hence, neighbors are more likely to share common venues.

**Analysis of inter-friend distance and check-ins.** Among 856 users of **H_FQ**, there are 271 users who have friends. We compute the distance between every pair of friends in **H_FQ** and the Jaccard similarity of their check-in venues (excluding all home venues). The distance between two friends is defined as the Euclidean distance between their home locations. When a friend pair shares no venues in common, its Jaccard similarity score is 0. To avoid the many 0-value Jaccard similarity affecting our correlation, we filter away such pairs. We compute the Pearson correlation between distance and Jaccard similarity of check-in venues. A negative correlation value -0.14 has been observed suggesting that the similarity between two friends' check-in venues tends to decrease with inter-friend distance. As not all friends' home locations are known, this weak correlation result could also be caused by data sparsity and hence should be investigated further.

### *E. Correlation between User/Venue Activity Level and Venue Distance*

In our data, there are users or venues who are very active in performing check-ins or receiving check-ins. Our intuition is that if a user has many check-ins, he is likely to check-in on venues farther away. Similarly, a venue with many check-ins is likely to be visited by users who live farther away.

To verify the first intuition, we rank users in the **H_FQ** dataset by the number of check-ins they perform and group them using the equal size strategy. That is, every 50 users in the sorted order is assigned to a bin and their average of mean distance between check-in venues and each home location is computed. This results in an average mean distance for each of the 17 bins as shown in Figure 3. The Pearson correlation score between the number of check-ins performed by users and the average mean distance of check-in venues is 0.2587. This implies that active users are likely to visit far away venues.



(a)

(b)

Fig. 3: (a) Correlation between check-in popularity of users and the average mean distance; (b) Correlation between check-in popularity of venue and the average mean distance

Next, we perform the same analysis on venues to ascertain that popular venues by check-in count (e.g., airports, shopping malls, etc.) attract more people from farther away home locations as shown in Figure 3. The size of bin increases to 500 because of larger number of venues. The Pearson correlation score between the venue's check-in popularity and the average mean distance of their check-in users is 0.0707. This correlation is smaller implying that popular venues may still only attract users from nearby home locations, and/or some less popular venues may attract mostly users from far away home locations.

## IV. Prediction of Check-in Activities

### *A. Task Definitions*

We define two check-in activity prediction tasks to evaluate the different types of features. There are several ways to define the check-in activity prediction task. Among them are the following two closely related yet simple tasks which we have chosen to be our focus. They are:

- **Check-in Count Prediction**: In this problem formulation, we assume that there are no previous check-ins from the target user to a new venue, and we want to predict how many check ins the target user will make on the venue. The predicted number can be any number larger than or equal to 0. This is a *regression analysis* problem. Formally, given a set of users $U$ and a set of venues $X$, and their check-in tuples each consisting of three elements, $(u, x, c_{ux})$, where $u \in U$, $x \in X$, and $c_{ux}$ ($\geq 0$) denote the number of check-ins a user $u$ has performed on venue $x$. Our task is to predict how many check-ins a target user $u'$ ($\in U$) will perform on a venue $x'$ ($\in X$).

- **Check-In Prediction**: This is a *binary classification* problem. Given a set of users $U$, a set of venues $X$, and a training set of tuples $(u, x, i_{ux})$'s where user $u \in U$, venue $x \in X$, and indicator $i_{ux} = 1$ if $u$ performs at least one check-in on $x$ (i.e., $c_{ux} > 0$); and $i_{ux} = 0$ otherwise. Our task is to predict whether a target user $u'$ ($\in U$) visits a venue $x'$ ($\in X$).

There are several useful applications that can benefit from the above two tasks. The two prediction tasks are involved in the recommendation of new venues to a user. Check-in prediction is the binary version of check-in count prediction that focuses more on accurate prediction of venues than the ranking of venues. The same prediction tasks can also be used in target advertising applications. As both prediction tasks are closely related, it is reasonable to expect the same feature set to be used in both.

### B. Feature classification

For each user-venue pair $(u, x)$, we would like to develop models that can perform both prediction tasks using a set of features that can be derived for the user-venue pair. Based on the analysis results in Section III, we have derived six different feature types, namely: (a) *user features*, (b) *venue features*, (c) *user-venue features*, (d) *friend-venue features*, (e) *neighbor-venue features*, and (f) *user-venue complex features*.

- *User features* $\mathbb{UF}$: These are features related to the user $u$ only.
  - UF1: number of venues visited by $u$: $\sum_{x \in X} i_{ux}$.
  - UF2: number of check-ins performed by $u$: $\sum_{x \in X} c_{ux}$
- *Venue features* $\mathbb{VF}$: These are features related to the venue $x$ only.
  - VF1: number of users who perform check-ins at $x$: $\sum_{u \in U} i_{ux}$
  - VF2: number of check-ins performed on $x$: $\sum_{u \in U} c_{ux}$
- *User-Venue features* $\mathbb{UVF}$: This feature set covers the direct interaction between $u$ and $x$ as well as between $u$ and venues related to $x$ which could affect $u$'s decision to perform check-ins at $x$.
  - UVF1: number of check-ins performed by $u$ at venues of type identical to the type of $x$: $\sum_{x' \in X, type(x') = type(x)} c_{ux'}$
  - UVF2: Euclidean distance (in meters) between the home location of $u$ and $x$ denoted by $dist(u, x)$.
  - UVF3: number of check-ins $u$ performs on the neighbors of $x$: $\sum_{x' \in X, dist(x', x) < 100} c_{ux'}$. Note that we empirically define a venue's neighbor to be another venue less than 100 meters away.
  - UVF4: number of check-ins $u$ performs on neighbors of $x$ that are of the same type: $\sum_{x' \in X, type(x') = type(x), dist(x', x) < 100} c_{ux'}$
- *Friend-Venue features* $\mathbb{FVF}$: This feature set captures the effect of friends of user $u$ on his decision to perform check-ins on $x$ or venues related to $x$.
  - FVF1: number of check-ins of friends of $u$ on $x$: $\sum_{u' \in F(u)} c_{u'x}$ where $F(u)$ denotes the set of friends of $u$.
  - FVF2: number of check-ins of friends of $u$ on venues similar to $x$: $\sum_{u' \in F(u), x' \in X, type(x') = type(x)} c_{u'x'}$.
  - FVF3: number of check-ins of friends of $u$ on venues near to $x$: $\sum_{u' \in F(u), x' \in X, dist(x', x) < 100} c_{u'x'}$
  - FVF4: number of check-ins of friends of $u$ on venues near and similar to $x$: $\sum_{u' \in F(u), x' \in X, dist(x', x) < 100, type(x) = type(x')} c_{u'x'}$
- *Neighbor-Venue features* $\mathbb{NVF}$: This feature set captures the effect of neighbors on $u$ performing check-ins at venues related to $x$ or also $x$ itself on $u$'s check-ins on $x$.
  - NVF1: number of check-ins that the neighbors of $u$ on $x$: $\sum_{u' \in N(u)} c_{u'x}$
  - NVF2: number of check-ins that the neighbors of $u$ on venues similar to $x$: $\sum_{u' \in N(u), type(x') = type(x)} c_{u'x'}$
  - NVF3: number of check-ins that neighbors of $u$ on venues near to $x$: $\sum_{u' \in N(u), dist(x', x) < 100} c_{u'x'}$
  - NVF4: number of check-ins that neighbors of $u$ on venues near to and similar to $x$: $\sum_{u' \in N(u), dist(x', x) < 100, type(x') = type(x)} c_{u'x'}$
- *User-Venue complex features* $\mathbb{UVIF}$: The features in this group try to capture the intuition in Section III-E by combining some features together.
  - UVIF1: The product between the inverse of distance between $u$ and $x$, and number of venues visited by $u$: $\frac{1}{dist(u,v)} \cdot \sum_{x' \in X} i_{ux'}$.
  - UVIF2: The product between the inverse of distance from $u$ to $x$ and number of users visiting $x$: $\frac{1}{dist(u,v)} \cdot \sum_{u' \in U} i_{u'x}$.
  - UVIF3: The product between number of venues visited by $u$ and number of users visiting $x$: $\sum_{x' \in X} i_{u,x'} \cdot \sum_{u' \in U} i_{u',x}$. The intuition is that if $u$ is active in performing check-ins, and $x$ is a venue attracting check-ins, $u$ is more likely to perform check-ins on $x$.

For features that require the home locations of users (i.e., UVF2, FVF2, NVF1, NVF2, NVF3, NVF4, UVIF1, and

UVIF2), we could replace the user home locations by center of the mass as defined in Section III-B. If features using center of mass could perform as well as those using home location, they will permit the check-in activity prediction task to be applied to users without home locations. Such users constitute more than 98% of all users as shown in Table I. For these features, we use subscript $h$ and $c$ in the feature name (e.g., UVF2$_h$ and UVF2$_c$) to denote the use of home location and center of the mass respectively. For example, UVF2$_h$ denotes the distance the user's home location to the venue $x$, while UVF2$_c$ denotes the distance from user's center of the mass to the venue $x$.

### *C. Prediction Methods*

To give us insights into the importance of features, we apply two methods for each task and measure the weight of parameters.

**Check-in count prediction.** We apply two methods, namely *linear regression* and *Support Vector Regression (SVR)*, to predict the number of check-ins by the target user and on the target venue.

The linear regression model for solving check-in count prediction is formulated as:

$$\begin{aligned} h_\Theta(u,x) &= \theta_0 + \theta_1 \cdot f_{ux_1} + \theta_2 \cdot f_{ux_2} + ... + \theta_n \cdot f_{ux_n} \\ &= F_{ux} \cdot \Theta \end{aligned} \tag{1}$$

$h_\Theta$ is the model we want to learn. $\theta_i$'s are the coefficients of the features $f_{ux_i}$'s of the user $u$ and venue $x$. $F_{ux}$ and $\Theta$ are the feature vector and $\theta_i$ vector respectively (with $f_{ux_0} = 1$). Our objective is to learn the parameters $\Theta$ that minimize the least square error of the predicted value and the actual value in the training set [2]. The optimum value of $\Theta$ which minimizes least square error can be derived using a closed form solution.

While linear regression minimizes least square error, SVR [20] aims to get the *flatness* (i.e., seeking a small $\Theta$) in Equation 1. Formally, we can rewrite it as a convex optimization problem with slack variables $\xi_{ux}$ and $\xi^*_{ux}$ for each user-venue pair $(u$,$x)$.

$$\begin{aligned} \underset{\Theta,\boldsymbol{\xi},\boldsymbol{\xi}^*}{\text{minimize}} \quad & \frac{1}{2}\|\Theta\|^2 + C\sum_{ux}(\xi_{ux} + \xi^*_{ux}) \\ \text{subject to} \quad & c_{ux} - F_{ux} \cdot \Theta \le \epsilon + \xi_{ux},\ \forall(u,x) \\ & F_{ux} \cdot \Theta - c_{ux} \le \epsilon + \xi^*_{ux},\ \forall(u,x) \end{aligned}$$

where $C > 0$ and $\epsilon$ are parameters to be learnt. $C$ is the trade off between getting *flatness* and the amount up to which deviations larger than $\epsilon$ are tolerated. This problem could be converted to its dual form for solving and $\Theta$ is reconstructed from the solution of the dual problem. In our experiments, we use the SVR implementation in the LIBSVM library [3].

**Check-in prediction.** We use two prediction methods for this task, namely *logistic regression* and *Support Vector Machine(SVM)*.

In logistic regression, the probability of check-in by user $u$ on venue $v$ is defined as a sigmoid function

$$P(i_{ux} = 1|\Phi) = \frac{1}{1 + \exp(-g_\Phi(u,x))} \tag{2}$$

where $\Phi$ is the parameter set to be learned from training data. $g_\Phi(u,x)$ is the linear combination between features of user $u$ and venue $x$ with parameter $\Phi$. We learn $\Phi$ by maximizing the log likelihood of training data with *L2-regularization* [15]. In our experiments, we use the logistic regression model provided by the LIBLINEAR library [8].

*Support Vector Machine(SVM)* [2] is the method to find a good hyperplane which has the largest distance to separate the different classes of training instances. Formally, it is the optimization problem with slack variable $\xi_{ux}$ for every pairs of user $u$ and venue $x$.

$$\begin{aligned} \underset{\Theta,\boldsymbol{\xi}}{\text{minimize}} \quad & \frac{1}{2}\|\Theta\|^2 + C\sum_{ux}\xi_{ux} \\ \text{subject to} \quad & i_{ux} g_\Phi(u,x) \ge 1 - \xi_{ux},\ \forall(u,x) \\ & \xi_{ux} \ge 0,\ \forall(u,x) \end{aligned}$$

where $C > 0$ is the regularization parameter. Similar to *SVR*, *SVM* can also be written as dual form and $\Theta$ can be reconstructed from the dual solution. In our experiments, we use the SVM implementation in the LIBSVM library [3].

In all our prediction methods, features are normalized by z-normalization technique [16] in all the data instances before they are used in training and testing.

## V. Experiments and Results

In this section, we describe the evaluation of our proposed methods for the *check-in count prediction* and *check-in prediction* tasks. The objective of the experiments is to determine: (a) the accuracy using different methods with different feature sets; (b) the importance of home location information; and (c) the importance of features in the two tasks.

### *A. Training and Test Data*

Based on **H_FQ**, we first constructed an experiment dataset with balanced user-venue pairs with non-zero check-in counts, and user-venue pairs with zero check-in counts. This dataset is used in both the prediction tasks. In the case of check-in prediction task, the check-in counts are converted into true (if $> 0$) and false (otherwise) before they are used.

The above dataset is created by selecting for each user-venue pair with non-zero check-ins another user-venue pair without check-ins. We do this by pairing the user with a randomly selected venue which the user has not performed check-in. This leads to a balanced set of positive and negatives instances. We finally derive an experiment dataset with 28,298 positive user-venue pairs and 28,298 negative user-venue pairs.

We divide above data into ten folds randomly such that each fold maintains a balanced set of positive and negative user-venue pairs. Each fold has between 632 and 645 users, between 4,723 and 4,817 venues, and around 5,660 user-venue pairs. To measure prediction accuracy, we use each fold of data for testing and the remaining nine folds for training, and average the accuracy across different choices of test fold. In this setup, it turns out that not every user/venue in each test fold can also be found in the remaining folds. When this arises, the test user-venue pairs concerned will be excluded from evaluation.

### *B. Experiment Setup*

**Baseline methods.** In addition to linear regression and support vector regression for check-in count prediction, we introduce three baseline methods for comparison.

- *Average check-in count* $B_1$: This baseline ignores the user and venue information, and returns the average check-in count of all user-venue pairs in the training data (both positive and negative). It therefore returns the same predicted value for all test user-venue pairs. We expect this baseline to perform poorly.
- *Average user's check-in count* $B_2$: This baseline returns the average check-in count of the user in the training data as the prediction. $B_2$ leverages on the previous check-ins of the user but not the venue.
- *Average venue's check-in count* $B_3$: This baseline returns the average check-in count of the venue in the training data. This is similar to $B_2$ except the choice of venue.

The three baseline methods can also be adapted for the check-in prediction task. $B_1$ is similar to a random guess which should yield an accuracy of 50%. For $B_2$ and $B_3$, we rank the test user-venue pairs by the decreasing predicted check-in count and select the top 50% pairs as the predicted pairs with some check-in.

**Feature Configurations.** For the four supervised methods, we adopt three feature configurations depending on the type of user location, i.e., home location and center of the mass. The configurations $G_h$ and $G_c$ include features with home location and center of the mass as user location respectively. The feature configuration $G_t$ includes all features, i.e., $G_t = G_h \cup G_c$.

**Performance measure.** We use *mean average square error* and *mean accuracy* to measure the performance of the check-in count prediction and check-in prediction respectively. The former is derived by averaging the squared differences between predicted and observed check-in counts for all the test user-venue pairs in each test fold, followed by taking the mean of the average squared errors across all test folds. Similarly for check-in prediction, we take the mean of the accuracy of prediction across all test folds.

### *C. Results of Check-in Count Prediction*

We first examine the performance of linear regression (LiR) and SVR. For SVR, we use the linear kernel and empirically set the parameters $\epsilon$ and $C$ to be 2 and 1 respectively. As our output variable is check-in count, $\epsilon = 2$ is a reasonable margin of tolerance for false prediction.

The check-in count prediction accuracy results are shown in Table III. The table shows that linear regression yield the best accuracy in feature configurations $G_h$ and $G_t$. SVR also performs better with $G_h$ and $G_t$ feature configurations. In other words, home location plays an important part improving the prediction accuracy. The feature configuration $G_c$ using center of the mass apparently yields less accurate results in both LiR and SVR methods. What is surprising is that the improvements of LiR and SVR over $B_2$ are quite small. Recall that $B_2$ returns user's average check-in count which actually works reasonably well for this prediction task. This result is interesting and will be further investigated in our future work.

TABLE III: Check-In Count Prediction Accuracy (Mean average square error)

| Methods/Features | $G_c$ | $G_h$ | $G_t$ |
|---|---|---|---|
| LiR | 4.29 | 4.03 | 4.03 |
| SVR | 4.33 | 4.24 | 4.23 |
| $B_1$ | 4.31 | | |
| $B_2$ | 4.26 | | |
| $B_3$ | 4.95 | | |

### *D. Result of Check-in Prediction*

Table IV shows the average accuracy of logistic regression (LoR) and SVM with three feature configurations $G_t$, $G_h$ and $G_c$. The results show that the two supervised methods outperform the baselines quite significantly with at least 10 percentile point difference. $B_2$ is still the strongest baseline but its accuracy is no longer close to the supervised methods. Among the feature configurations, $G_t$ and $G_h$ again perform better than $G_c$ confirming that center of mass is inferior than home location in this prediction task. The findings are consistent with that in check-in count prediction.

TABLE IV: Mean Accuracy of Check-In Prediction methods

| Method | $G_c$ | $G_h$ | $G_t$ |
|---|---|---|---|
| LoR | 77.54% | 78.34% | 78.78% |
| SVM | 77.88% | 78.02% | 78.34% |
| $B_1$ | 50% | | |
| $B_2$ | 67.47% | | |
| $B_3$ | 56.23% | | |

## VI. Feature Analysis

As the earlier results show that the feature configuration $G_t$ gives the best prediction results, we want to further analyze the importance of features. Table V shows the coefficient learnt for each feature in check-in count prediction and check-in prediction with the $G_t$ configuration. The coefficients with larger magnitudes are deem to be more important and thus shown in boldface.

The Spearman correlation between the feature coefficients of linear regression and SVR is 0.8216 while the Spearman correlation between that of SVM and logistic regression is 0.7165. This shows us that the different methods have positive correlation in ranking the feature importance.

For the check-in count prediction task, the features VF1, VF2, UVF4, and UVIF1$_h$ have the largest absolute coefficients returned by both linear regression and SVR. In particular, UVIF1$_h$ and VF2 are the more important features. The former is related to the combination of user's level of active check-ins and the distance between his home location and venue. The latter is related to venue popularity by check-in count. For the check-in prediction task, logistic regression and SVM share three out of the four most important features, namely VF1, UVF3 and UVIF3. VF1 is related to popularity of venue by user count, UVF3 is related to the familiarity of the venue's neighborhood through check-ins, and UVIF3 is related to the combination of user's level of active check-ins and the popularity of venue by user count.

Interestingly, both friend-venue feature set ($\mathbb{FVF}$) and neighbor-venue feature set $\mathbb{NVF}$ do not contribute much to the accuracy of the two prediction tasks. This can be attributed to many friends or neighbors of users not found in our **H_FQ** dataset.

TABLE V: Average coefficients of features.

| | | Count Prediction | | Check-in Prediction | |
|---|---|---|---|---|---|
| | | LiR | SVR | LoR | SVM |
| User features | UF1 | -1.66 | -0.02 | 4.73 | **0.58** |
| | UF2 | 4.57 | 0.04 | -0.85 | -0.14 |
| Venue features | VF1 | **-14.49** | **-0.15** | **18.36** | **1.89** |
| | VF2 | **47.17** | **0.21** | 1.66 | -0.33 |
| User-Venue features | UVF1 | 2.30 | 0.04 | 4.45 | 0.21 |
| | UVF2$_h$ | -1.04 | -0.06 | -2.25 | -0.26 |
| | UVF2$_c$ | -0.96 | -0.02 | -3.59 | -0.21 |
| | UVF3 | 7.19 | 0.10 | **18.72** | **3.73** |
| | UVF4 | **21.75** | **0.15** | 2.59 | 0.45 |
| Friend-Venue features | FVF1 | 5.30 | 0.03 | 0.80 | 0.07 |
| | FVF2 | 0.51 | 0.00 | 0.64 | -0.01 |
| | FVF3 | 0.40 | -0.00 | 1.51 | 0.00 |
| | FVF4 | -5.59 | -0.00 | -0.20 | -0.01 |
| Neighbor-Venue features | NVF1$_h$ | -0.83 | 0.04 | 1.04 | 0.10 |
| | NVF1$_c$ | -3.19 | 0.00 | -0.12 | 0.00 |
| | NVF2$_h$ | -1.31 | -0.04 | -2.22 | -0.12 |
| | NVF2$_c$ | -0.73 | -0.00 | -0.78 | -0.02 |
| | NVF3$_h$ | 0.29 | -0.00 | 3.68 | 0.06 |
| | NVF3$_c$ | -1.37 | -0.01 | -0.17 | -0.02 |
| | NVF4$_h$ | 0.04 | 0.01 | 0.19 | -0.01 |
| | NVF4$_c$ | -0.38 | 0.00 | 0.01 | -0.01 |
| User-Venue complex features | UVIF1$_h$ | **47.17** | **0.49** | 1.66 | -0.03 |
| | UVIF1$_c$ | 1.50 | 0.02 | 0.74 | 0.01 |
| | UVIF2$_h$ | 4.22 | 0.05 | **7.90** | 0.29 |
| | UVIF2$_c$ | -3.53 | -0.00 | 0.42 | 0.01 |
| | UVIF3 | 1.59 | 0.11 | **7.47** | **3.48** |

## VII. Conclusion and Future Works

In this paper, we studied the check-in patterns of users through their exact home locations. To the best of our knowledge, this is the first work using the exact home location of users to analyse and predict the check-in behavior of users in location-based social networks. Our empirical analysis shows that users are more likely to perform check-ins on places near their home locations. As a result, neighbors are more likely to share more common check-in venues. We also found active users tend to perform check-ins on places farther away from their home locations.

Due to the setup of the problem, we show that social and neighbor relationships are not as strong as the distance of venue from the user's home location and venue popularity in influencing user check-in decisions. Our experiment also shows that the supervised methods in general can predict the check-in count and check-in decisions more accurately than the baseline methods.

In the future, we plan to extend our work to study more detailed check-in behavior using additional dimensions of information. We plan to include temporal information (e.g., weekday versus weekend, hour of the day, etc.) of check-ins as well as user attributes in the analysis and prediction tasks. Check-ins could also be contributed by events and detecting events beyond the user's usual check-in patterns is also an interesting research direction.

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

## References

[1] R. Assam and T. Seidl. Check-in location prediction using wavelets and conditional random fields. In *ICDM*, 2014.

[2] C. M. Bishop et al. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

[3] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM TIST*, 2, 2011.

[4] J. Chang and E. Sun. Location 3: How users share and respond to location-based data on social networking sites. In *ICWSM*, 2011.

[5] C. Cheng, H. Yang, I. King, and M. R. Lyu. Fused matrix factorization with geographical and social influence in location-based social networks. In *AAAI*, volume 12, 2012.

[6] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: User movement in location-based social networks. KDD, 2011.

[7] H. Cramer, M. Rost, and L. E. Holmquist. Performing a check-in: emerging practices, norms and'conflicts' in location-sharing using foursquare. In *International Conference on Human Computer Interaction with Mobile Devices and Services*, 2011.

[8] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *JMLR*, 9, 2008.

[9] H. Gao, J. Tang, and H. Liu. gscorr: modeling geo-social correlations for new check-ins on location-based social networks. In *CIKM*, 2012.

[10] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. Trajectory pattern mining. In *KDD*, 2007.

[11] J. Krumm. Inference attacks on location tracks. In *PERVASIVE*, 2007.

[12] J. Krumm and E. Horvitz. Predestination: Inferring destinations from partial trajectories. In *UbiComp*. 2006.

[13] Z. Li, B. Ding, J. Han, R. Kays, and P. Nye. Mining periodic behaviors for moving objects. In *KDD*, 2010.

[14] D. Lian, V. W. Zheng, and X. Xie. Collaborative filtering meets next check-in location prediction. In *WWW*, 2013.

[15] C.-J. Lin, R. C. Weng, and S. S. Keerthi. Trust region newton method for logistic regression. *JMLR*, 9, 2008.

[16] M. L. Marx and R. J. Larsen. *Introduction to mathematical statistics and its applications*. Pearson/Prentice Hall, 2006.

[17] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. Mining user mobility features for next place prediction in location-based services. In *ICDM*, 2012.

[18] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An empirical study of geographic user activity patterns in foursquare. *ICWSM*, 11, 2011.

[19] D. Preoţiuc-Pietro and T. Cohn. Mining user behaviours: a study of check-in patterns in location based social networks. In *ACM Web Science*, 2013.

[20] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3), 2004.

[21] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968), 2010.

[22] M. A. Vasconcelos, S. Ricci, J. Almeida, F. Benevenuto, and V. Almeida. Tips, dones and todos: uncovering user profiles in foursquare. In *WSDM*, 2012.

[23] J. Ye, Z. Zhu, and H. Cheng. What's your next move: User activity prediction in location-based social networks. In *SDM*, 2013.

[24] Y. Ye, Y. Zheng, Y. Chen, J. Feng, and X. Xie. Mining individual life pattern based on location history. In *MDM*, 2009.