Analyzing Location Predictability on Location-Based Social Networks

Defu Lian[1,2], Yin Zhu[3], Xing Xie[2] and Enhong Chen[1]

[1]University of Science and Technology of China
[2]Microsoft Research
[3]Hong Kong University of Science and Technology
liandefu@mail.ustc.edu.cn, yinz@cse.ust.hk, xingx@microsoft.com
cheneh@ustc.edu.cn

Abstract. With the growing popularity of location-based social networks, vast amount of user check-in histories have been accumulated. Based on such historical data, predicting a user's next check-in place is of much interest recently. There is, however, little study on the limit of predictability of this task and its correlation with users' demographics. These studies can give deeper insight to the prediction task and bring valuable insights to the design of new prediction algorithms. In this paper, we carry out a thorough study on the limit of check-in location predictability, i.e., to what extent the next locations are predictable, in the presence of special properties of check-in traces. Specifically, we begin with estimating the entropy of an individual check-in trace and then leverage Fano's inequality to transform it to predictability. Extensive analysis has then been performed on two large-scale check-in datasets from Jiepang and Gowalla with 36M and 6M check-ins, respectively. As a result, we find 25% and 38% potential predictability respectively. Finally, the correlation analysis between predictability and users' demographics has been performed. The results show that the demographics, such as gender and age, are significantly correlated with location predictability.

Keywords: Location predictability, entropy, LBSN

1 Introduction

With the proliferation of smart phones and the development of positioning technologies, users can obtain location information more easily than ever before. This development has triggered a new kind of social network service - location-based social networks (LBSNs). In a LBSN, people can not only track and share location-related information of an individual, but also leverage collaborative social knowledge learned from them. "Check-in" is such user-generated and location-related information, being used to represent the process of announcing and sharing users' current locations in LBSNs.

In this paper, we are interested in predicting users' future check-in locations based on their location histories accumulated in LBSNs. In particular, we attempt to determine at which Point Of Interest (POI), such as a clothing store or

a western restaurant, a user will check in next. Though this problem has been recently investigated [1,2,3], there is little study on the limit of predictability, i.e., to what degree the next check-in locations are predictable, and its correlation with users' demographics. We believe this study will bring valuable insights to the design of prediction algorithms and help to understand users' behavior from both social and physical perspectives.

The limit of location predictability was first studied on cell tower location sequences in [4]. The authors discovered a 93% potential predictability in human mobility. However, a check-in trace is quite different from a cell tower trace in the following three aspects. First, check-in is a proactive behavior comparing to the passive recording of cell tower traces. In other words, a user might not check in at boring places where he has actually been but may check in at locations where there is no visiting behavior. Therefore, check-in locations are usually discontinuous, and many important mobility patterns could have been lost. Second, the spatial granularity of check-in locations is much finer than cell tower locations (e.g., a point location versus an area of one square kilometers). Thus there are more candidate locations to choose for check-in so that it is much more difficult to predict next check-in location. Last but not least, users are equipped with rich profile information and social relationships, since their check-ins are usually shared on different social networks. This would be helpful for developing more accurate algorithms. In our work, we analyze the problem of check-in location prediction in the presence of these characteristics.

To study the limit of location predictability, we begin with estimating the entropy of an individual check-in trace by first considering an individual check-in trace as a sample of underlying stochastic processes and then calculating the entropy of stochastic processes. We then leverage Fano's inequality [5] to transform the estimated entropy into the limit of predictability for each user. The limit of check-in location predictability is measured for each user on two large-scale check-in datasets from Jiepang and Gowalla with 36M and 6M check-ins, respectively. As a result, we find 25% and 38% potential predictability on these two datasets, respectively.

However, according to our observation, the variance of location predictability among population is large. It implies there is large diversity of human mobility patterns among population. To better understand such large diversity, we can study the difference in predictability of users with different demographics. Particularly, we will perform correlation analysis between predictability and demographics. This task can be more easily done than ever before since users have been already equipped with rich profile information on social networks, including gender, age, social relationship and so on. By conducting case studies on these check-in datasets, we show that the demographics including users' gender, age and influence (measured as the number of followers) as well as the repetitiveness of check-ins (measured as the ratio of the number of check-ins to locations) are significantly correlated with location predictability. More specifically, the mobility of students is higher predictable since their activity areas are usually constrained around the campus and their mobility patterns tend to be more reg-

ular; the users with high social influence are hard to predict because they don't usually repeat to check in at those familiar locations. In this case, it is evident that incorporating demographics into the prediction task could be beneficial.

2 Related Work

The limit of location predictability was first studied in [4] on cell tower data, on which they derived an upper bound of predictability from the entropy of the individual location sequence and found a 93% potential predictability in human mobility. They also studied on a lower bound, Regularity, which predicted the next location as the most frequently visited location at given hour of week. Following them, in [6], the authors studied the predictability from mobile sensor data and also found a high potential predictability on mobile sensor data. And in [7] the authors investigated the scaling effects on the predictability using the high resolution GPS trajectories and derived another equivalent statistical quantities to the predictability. In their conclusion, they stated that high predictability was still present at the very high spatial/temporal resolutions. Although all these work focused on the analysis of the limit of predictability, their mobility data differs from ours in the following two major aspects. First, the check-in trace is more discontinuous since a user might not check in many places which he has actually visited. Second, the spatial granularity of check-ins is even finer than physical locations since there might be many different semantic locations in the same physical location. These properties lower the check-in location predictability and only achieve a 25%-40% potential predictability.

As for the correlation between mobility patterns and demographics, in [8], the authors analyzed mobility patterns based on the travel diaries of hundreds of volunteers and figured out people with different occupation had distinct mobility patterns. In particular, students and employees were more tending to move among those frequented locations than retirees. From the perspective of predictability, it seems that students and employees were more easily predicted. The direct correlation between predictability and users' demographics was also studied in [4] on cell tower traces logged by hundreds of volunteers with some demographics. They concluded that there were no significant gender- or age-based differences. Different from theirs, we perform analysis on check-in traces of hundreds of thousands of users on social networks, where users are usually equipped with rich profile information. The results of analysis show that user's demographics including age, gender, social influence and so on, are significantly correlated with location predictability.

3 Check-in Datasets

We perform our analysis on two check-in datasets. The first check-in dataset is from Jiepang, which is a Chinese location-based social network similar to Foursquare. For the sake of protecting privacy, in these LBSNs, users' historical check-ins are not shown to strangers. Thus we cannot directly obtain users'

check-ins from these LBSNs without becoming their friends. However, users may share their check-ins as tweets on other social networking platforms, such as Weibo and Twitter. For example, Jiepang check-ins are synchronized on Weibo as a particular type of tweets (called location tweets). Thus these check-ins can be crawled from these social networking platforms via their open APIs. Some check-in datasets were also crawled in this way [2,3].

We crawled 36,143,085 Jiepang check-ins at 1,000,457 POIs from 454,375 users via the Weibo API from March. 2011 to March. 2013, where each user has 80 check-ins and check in at 47 POIs on average. If we distribute these check-ins into their date, we find that each user only make 1.5 check-ins each day on average. If we distribute these POIs into 3 km^2 regions, each region owns 13 POIs on average and up to 13,068 POIs in the maximal case. In addition, users on Weibo may fill their profile information more precisely so we also crawled these data, including age, gender, and social relationship as well as tags.

The other check-in dataset, used in [9] and crawled from Gowalla from Feb. 2009 to Oct. 2010, contains 6,423,854 check-ins at 1,280,969 POIs from 107,092 users, where each user has 60 check-ins and check in at 37 POIs on average. If we distribute these check-ins into their date, we find that each user only make 2.1 check-ins each day on average. If we distribute these POIs into 3 km^2 regions, each region owns 7 POIs on average and up to 3,940 POIs in the maximal case.

Based on the above statistics, it is easily observed that the frequency of check-ins is significantly smaller than calling or messaging (SMS) and that location density on check-in datasets is significantly higher than cell towers since each cell tower covers a 3-km^2 perception area on average.

In order to guarantee that the entropy of location sequence is well estimated, we only reserve those users with more than 50 check-ins. As a result, 144,053 and 27,693 users are then kept on Jiepang and Gowalla, respectively. All remaining users on Jiepang have gender information while 53,377 out of them have age information. Moreover, they have 3.9 tags and 15 followees on average. Based on the filtered datasets, we perform extensive analysis after presenting the limits of predictability and then compare them with cell tower traces.

4 Location Predictability

Assume we predict the n^{th} check-in location L_n for user u, given her past location sequence of length $n-1$, $h_{n-1} = \{l_1, l_2, ..., l_{n-1}\}$. From the probabilistic perspective, we need to model the probability distribution of L_n given h_{n-1}, i.e., $P(L_n|h_{n-1})$. In the context of prediction, we choose the location $\hat{l}$ with the maximum probability

$$\hat{l} = \arg\max_{l} P(L_n = l|h_{n-1}). \tag{1}$$

Intuitively, if the distribution of $P(L_n|h_{n-1})$ is flat, the prediction $\hat{l}$ with the maximum probability has a low likelihood of being correct; if the distribution peaks at location $\hat{l}$ significantly, then the prediction can be made with high confidence. Thus the probability at $\hat{l}$ (denoted as $\pi(h_{n-1})$) contains the full predictive

power including the potential long range dependency. Summing $\pi(h_{n-1})$ over all possible sequences of length $n-1$, the predictability at the n^{th} location is defined as

$$\Pi(n) \equiv \sum_{h_{n-1}} P(h_{n-1})\pi(h_{n-1}), \tag{2}$$

where $P(h_{n-1})$ is the probability of observing h_{n-1}.

After averaging the predictability over all time indices and taking the limit, each user's predictability Π is defined as

$$\Pi \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Pi(i) \tag{3}$$

The limit that Π can reach is estimated by first calculating the entropy of a user's check-in location sequence using non-parametric approaches, and then transforms the estimated entropy into the limit of Π using Fano's inequality [5].

4.1 Entropy of Check-in Location Sequence

The history of check-in locations of a user can be considered as one sample path of its underlying stochastic process, e.g., Markov process. Therefore, the entropy estimation of the location history is equivalent to deriving the entropy rate of the stochastic process. According to the definition of entropy, the entropy rate of a stationary stochastic process $\mathcal{L} = \{L_i\}$ is defined as,

$$S \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} S(L_i|H_{i-1}) \tag{4}$$

where $S(L_i|H_{i-1})$ is the conditional entropy of L_i given H_{i-1}, which is a random variable corresponding to h_{i-1} (i.e., the past location sequence of length $i-1$). If the stochastic process lacks any long range temporal correlations, i.e., $P(L_i|h_{i-1}) = P(L_i)$, its entropy is $S^{unc} = -\sum_{l=1}^{N} P(l)\log_2 P(l)$, where $P(l)$ is the probability of being at location l and N is the number of locations. In this case, the user moves around N locations according to previous visiting frequency, which we named as the MostFreq algorithm. Another special entropy of interest is the random entropy $S^{rand} = \log_2 N$, obtained when $P(l) = \frac{1}{N}$. In this case, the user moves around N locations randomly, which we named as the Random algorithm. It is obvious that $0 \le S \le S^{unc} \le S^{rand} < \infty$.

One practical way of calculating the entropy of the user's location history is to fix a underlying stochastic process model and then estimate its parameters, e.g., transition probability of first-order Markov process, and finally derive the entropy rate. This method follows a parametric way and somewhat over-specific. From the non-parametric perspective it can also be achieved to use an estimator based on Lempel-Ziv data compression [10]. This method doesn't assume the stochastic process model and thus is more general. In [10], the authors discussed three kinds of LZ estimators and proved that they can converge to the real

entropy of a time series when the length of observation sequence is approaching infinity. They applied them to calculate the entropy of English texts (number of bits storage). One estimator for a time series with n steps is defined as follows:

$$S \approx \frac{\ln n}{\frac{1}{n}\sum_{i=1}^{n} \Lambda_i^i} \tag{5}$$

where Λ_i^i is the length of the shortest substring starting at position i which doesn't previously appear from position 1 to $i-1$.

To get the real entropy for any user, her location must be recorded continuously (e.g., hourly). However, the cell tower traces used in [4] only contain locations when a person uses her phone, e.g., she sends a short text message or makes a call, and thus exhibits discontinuity and bursting characteristics in temporal dimension. To handle bursting, the authors merged locations within the same hour. To deal with discontinuity, they first studied the relationship between the entropy of discontinuous location history and the degree of discontinuity, and then extrapolated the entropy where the degree of discontinuity was zero. However, in addition to discontinuity and bursting, check-ins are at the granularity of POIs instead of regions in the cell tower traces. POIs are physical coordinates with semantic labels so that it is possible for different POIs to share the same physical coordinates. Thus POIs are even finer-grained than physical coordinates. For instance, shops in the same building share the physical location. As the check-in POIs within the same hour may have different semantic labels, it is difficult to merge them so that the subsequent extrapolation cannot be applied. Instead, we can simply use the entropy calculated from the check-in history. This is reasonable to some extent since the benefit of extrapolation results from imputing unseen locations while imputing unseen check-in location is more difficult than imputing physical locations.

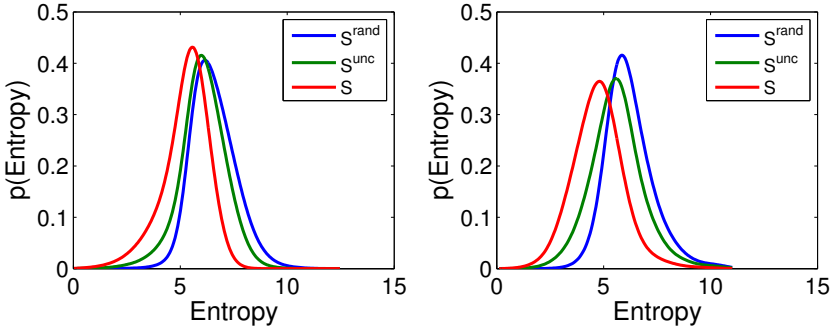

Fig. 1. The distribution of S, S^{unc} and S^{rand} across user population on Jiepang (left) and Gowalla (right).

Next we measure three entropy quantities S, S^{unc} and S^{rand} for each user, and show their probability distribution across users on Jiepang and Gowalla separately in Figure 1. Compared to the results obtained from cell tower traces,

there is no big gap between $P(S)$ and $P(S^{rand})$ on both check-in datasets. Indeed, $P(S^{rand})$ peaks around 5.8 on both check-in datasets, indicating if we assume users moved randomly, their next locations can be found on average in any of $2^{5.8}$=56 locations. However, $P(S)$ peaks around 5.59 and 4.83 on Jiepang and Gowalla, respectively. In other words, if we make a prediction with the help of their past history, we reduce less than 2 bits of uncertainty and must choose among $2^{5.59}$=48 and $2^{4.83}$=28 locations on average, respectively, which is much larger than the corresponding number ($2^{0.8}$=1.74) obtained from cell tower traces. Therefore, the prediction of check-in location is more difficult than the cell tower location. To get deep understanding on the difficulty of prediction on LBSNs, we compile statistics on what percentage of transition across locations will repeat. The result indicates that there are only 3.4% and 6.5% repetitive transitions across locations on Jiepang and Gowalla, respectively. This may be because users' proactive check-in behaviour renders checking in at locations without actual visit and missing check-ins at locations where they often go. To continue analyzing Figure 1, we observe that the difference between $P(S^{rand})$ and $P(S)$ on the Jiepang check-in dataset is smaller than on the Gowalla check-in dataset, which means that check-in location prediction on Jiepang is more difficult. This is in line with the previous results that there are larger repetitive transitions across locations on Gowalla than on Jiepang. Comparing $P(S^{unc})$ with $P(S^{rand})$, there is only a small gap on the Jiepang check-in dataset, which indicates that a large number of locations are checked in only once since the average times of users' check-in at POIs is less than 2. The extra part of $P(S)$ over $P(S^{unc})$ can be explained by the temporal correlation between locations in the location sequence and thus helps us to understand the effect of the sequential patterns. Due to their small gap, we could foresee the limited benefit from sequential patterns in the check-in traces.

4.2 Limit Analysis and Discussions

As soon as we get three quantities of entropy, we can transform them to the limit of their corresponding prediction algorithms. For the aforementioned predictability Π, it satisfies Fano's inequality. That is, if a user with entropy S moves between N locations, her predictability Π meets this condition $\Pi \leq \Pi^{max}$, where Π^{max} is the root of following equation

$$S = S_F(\Pi^{max}) \tag{6}$$

$S_F(p) = (1-p)\log_2(N-1) + H(p)$ is a Fano function ($H(p)$ is a binary entropy), which is concave and monotonically decreases with p when $p \in [\frac{1}{N}, 1)$. Therefore, the satisfaction of $\Pi \leq \Pi^{max}$ only requires $S_F(\Pi) \geq S = S_F(\Pi^{max})$ since it is easily verified that $\Pi \geq \frac{1}{N}$ and $\Pi^{max} \geq \frac{1}{N}$. Based on the concavity and monotonicity properties of $S_F(p)$ as well as Jensen's inequality, $S_F(\Pi) \geq S$ is equivalent to $S_F(\pi(h_{n-1})) \geq S(L_n|h_{n-1})$. The latter inequality is simply the well-known Fano's inequality when the probability of error prediction is $1 - \pi(h_{n-1})$ so that $\Pi \leq \Pi^{max}$ is proved.

Since $\Pi^{max} \geq \frac{1}{N}$, Eq (6) has a unique root. We can leverage any root finding algorithm, e.g., Newton's method to find the solution of Π^{max}. Similarly from S^{rand} and S^{unc}, we can determine Π^{rand} and Π^{unc}, which are limits of Random and MostFreq, respectively.

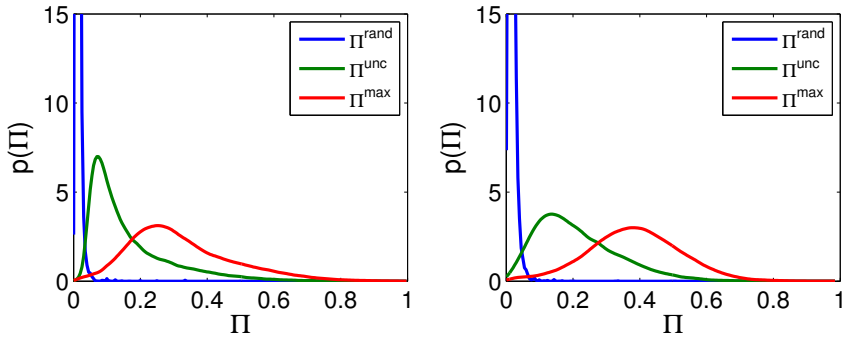

Fig. 2. The distribution of Π^{max}, Π^{unc}, Π^{rand} on Jiepang (left) and Gowalla (right).

After determining Π^{max}, Π^{unc} and Π^{rand} for each user using Eq (6), we demonstrate their distribution over all users in Figure 2. We can see that Π^{max} on the Jiepang and Gowalla check-in datasets peaks around 0.25 and 0.38, respectively, which means that *no matter* how good the algorithm is, its accuracy can not be larger than 25% and 38%, respectively when using their individual check-in history. In addition, we can see that the variance of Π^{max} is larger than that obtained from cell tower traces and thus the predictability in terms of check-in location varies more from person to person. In other words, some people show high regularity while others do not. Π^{unc} shows a similar trend to Π but is left shifted and more narrowly distributed. Thus Π^{unc} shows more universal patterns across user population than Π^{max}. This observation agrees with the phenomena that users check in at the most locations only once.

Although there is 25%-38% potential predictability on the check-in data, there are two assumptions behind. The first is not to impute unseen check-in locations. The reason of making this assumption is that it is difficult to impute unseen check-in locations in practice. The second assumption is not to resort to other information, such as check-ins of friends and similar users. This assumption mainly results from the dominated effect of individual check-in history according to our observation and the conclusion in [1,2]. In the future, we can relax the second assumption to consider these information so as to get a higher limit of predictability. Although the potential predictability is much lower than that of cell tower traces, it is still a theoretically tight bound, which is difficult to achieve in practice. In order to approach this bound, according to the definition of entropy, we should leverage all orders of sequential patterns, from 0 order (MostFreq) to possibly highest (the number of all check-ins) order. In practice, significant high-order sequential patterns are difficult to discover due to limited check-in history, so the combination of 0 order and other low order sequential patterns

can be a possible way of designing prediction algorithm [3]. Additionally, for the sake of approaching this bound, according to Fano's inequality, the probability of error $1 - \pi(h_{n-1})$ should be distributed as uniformly as possible over all locations except the most possible one $\hat{l}$. To achieve this, we can introduce other information, such as the time of next check-in, to increase $\pi(h_{n-1})$.

5 Predictability and Demographic

After calculating the limit of location predictability based on individual check-in history, we further study the correlation between location predictability and several users' demographics. Such study is important since it could provide evidence for the demographic-based advertisement targeting and the demographic-boosted prediction task. In this study, the demographics are split into two categories: categorical and numerical. Categorical demographics include gender and age group. Numerical demographics include number of locations, number of check-ins, and number of followers. The age information is numerical in practice, but we quantized it into the following ordinal groups, i.e., "<19", "19-23", "24-28", "29-33", "34-38", ">38" for better visualization. Since the ages of most users are distributed within $[19, 38]$, users both younger than 19 and older than 38 are aggregated into separate groups.

For categorical variables, we perform analysis of variance (ANOVA) test for the statistical correlation between a demographic measure and user's predictability. For numerical variables, since we don't know the concrete form of their correlation, we calculate a non-parametric correlation, i.e., Kendall rank correlation coefficient, and perform non-parametric hypothesis test, i.e., tau test, to see the significance of their correlation.

Table 1. Analysis of variance (ANOVA) for testing the correlation between gender as well as age and predictability. F is F-statistics and p is the p-value of statistical test

	Gender		Age	
	F	p	F	p
Predictability	4584	<1e-10	260.1	<1e-10

Before performing ANOVA test, we first check the assumption of ANOVA test on the categorical variables and predictability. The result of testing shows the assumption can be satisfied with a p-value smaller than 1e-10. The result of ANOVA test is shown in Table 1. From this table, we observe that the correlation of predictability with the categorical demographics including gender and age group is significant. And gender is more correlated with predictability, which indicates that male users and female users show different degree of predictability. In order to see how these categorical variables are correlated with predictability, we draw the box plot of predictability with respect to these categorical variables and show them in Figure 3. From them, we have the following observations: 1) male users show higher regularity than female users. According to the statistics

of the categories of POIs, the four most frequent check-in categories from female users are residential, coffee shop, shopping mall and chaffy dish while the four most frequented check-in categories made by male users are residential, airport, office building and subway station. Thus this observation sounds reasonable. 2) young users (age<24) are easier to predict. This is because these young users are mainly students at school so that their check-ins are constrained around their campus. According to the statistics of the POI categories, three most frequent check-in categories by these young users are teaching building, dormitory and campus. According to the analysis to the tags of each user, 10% of these young users are tagged by "*students*" while only 1% of elder users are such tagged.

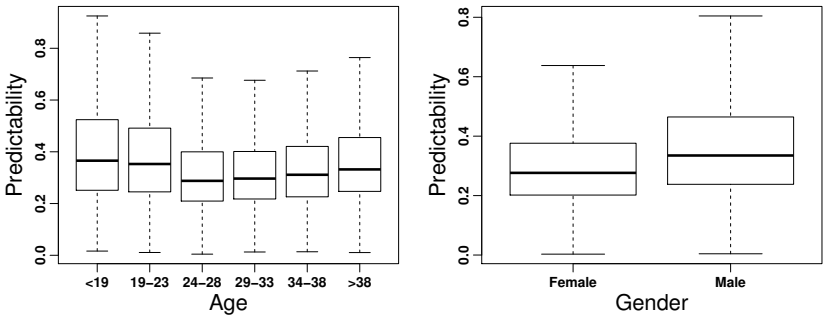

Fig. 3. Box plot of predictability with respect to gender and age

Next we study the correlation between numerical variables and predictability and show the results of tau test in Table 2. From this table, we can answer the following two questions. 1) Are users with larger social influence harder to predict? We measure social influence as the number of followers (#F) in this paper. Users with more followers are more cautious about their reputation so that they will not check in at those boring locations, such as home, subway station. According to this table (last column), the answer to this question is yes and thus the immediately preceding explanation is justified. 2) which of these three factors, the number of locations (#L), the number of check-ins (#C) and the ratio of the number of check-ins to the number of locations which we name as *CLR*, are consistently and significantly correlated to predictability? From this table, we see that the correlation between predictability and the number of locations is significant but not consistent on Jiepang (positively correlated) and Gowalla (negatively correlated). More discussions are provided in Figure 4 and in the subsequent paragraphs. As for the number of check-ins and its ratio *CLR* to the number of locations, they are both significantly correlated to the limit of location predictability on both datasets from the perspectives of statistics testing. However, *CLR* is more strongly correlated with location predictability than the number of locations. The principle reason is that the larger number of check-ins doesn't necessarily indicate more repetitive patterns since some users like to check in at many neighbour locations which they didn't visit in practice.

Table 2. Kendall rank correlation test between numerical profile variables and predictability. Z means the Z-statistics in τ test of Kendall rank correlation, τ is the correlation coefficient and p is the p-value. The cells with bold font indicate negative correlation and the cell with bold italic font shows insignificant correlation

	Dataset	Stat	#L	#C	CLR	#F
UB	Jiepang	Z	42.6	184.5	516.5	**-10.3**
		p	<1e-10	<1e-10	<1e-10	**<1e-10**
		τ	0.075	0.325	0.907	**-0.018**
	Gowalla	Z	**-56.7**	24.4	212.1	
		p	**<1e-10**	<1e-10	<1e-10	
		τ	**-0.228**	0.098	0.850	

To find how predictability covariates with the number of locations and *CLR*, we plot them together with predictability in Figure 4. We can see that the relationship between the limit of predictability and the number of location is really inconsistent on the two check-in datasets according to Figure 4(a). Specifically, when the number of locations is larger than 52, predictability of users from Jiepang is increasing while on Gowalla it is keeping comparatively stable. The reason behind may be that Jiepang users who check in at more locations may also check in more at their regular locations. To justify, we compute the correlation between the number of check-ins and the ratio of the number of check-ins to the number of locations (*CLR*). We find that on Jiepang there exists positive Kendall rank correlation ($\tau = 0.155$) between them while on Gowalla they are negatively correlated ($\tau = -0.208$). This means that when checking in at more locations, the average number of check-ins at locations is increasing on Jiepang. This implies that these users also check in more at these familiar locations. However, a contrast trend is observed on Gowalla. Thus, the correlation between predictability and the number of locations on check-ins seems incompatible with the result discovered in [4], that regularity was inversely proportional to $N^{-\frac{1}{4}}$, where N is the number of locations. However, according to Figure 4(b), the correlation between the limit of location predictability and *CLR* is consistently positive on both datasets. This indicates larger *CLR* could imply more repetitive patterns so that users' behaviour can be more accurately predicted.

6 Conclusion and Future Work

We have analyzed check-in location predictability on two large scale check-in datasets from Jiepang and Gowalla, and found 25% and 38% potential predictability respectively. Then we have studied the correlation between location predictability and users' demographics. The results show that the check-in behaviour of the male users and the young students are more higher predicted. By comparing the correlation between location predictability and the number of locations on two check-in datasets, we have not observed the universal correlation between them. In other words, the number of locations is not directly correlated to location predictability. However, the number of check-ins and its ratio to the

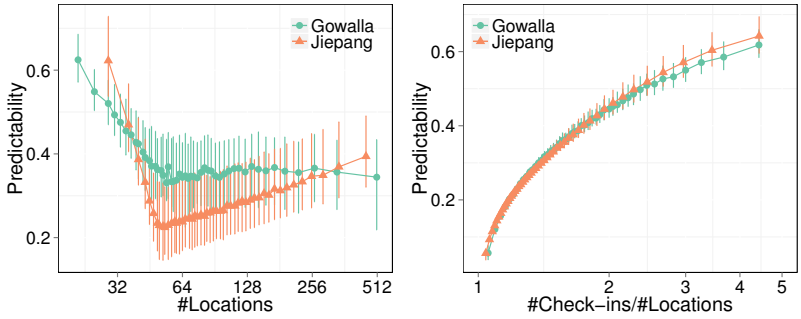

Fig. 4. Predictability with respect to some continuous profile variables. Continuous profile variables are quantized into 50 groups by its uniform quantiles. For each group, we show the median, 25% and 75% quantile of predictability in the error bar plot.

number of locations was significantly and positively correlated to predictability, but the degree of the ratio's correlation to predictability is stronger.

In the future, we will extend our predictability analysis from a single user to user groups and study location predictability in the presence of both real friendship on social network and virtual friendship based on location visiting history. This analysis can be helpful to predict check-ins at novel locations where users have never visited before.

References

1. Chang, J., Sun, E.: Location3: How users share and respond to location-based data on social. In: Proc. of ICWSM'11. (2011)
2. Noulas, A., Scellato, S., Lathia, N., Mascolo, C.: Mining user mobility features for next place prediction in location-based services. In: Proc. of ICDM'12, IEEE (2012) 1038–1043
3. Gao, H., Tang, J., Liu, H.: Exploring social-historical ties on location-based social networks. In: Proc. of ICWSM'12. (2012)
4. Song, C., Qu, Z., Blumm, N., Barabási, A.: Limits of predictability in human mobility. Science **327**(5968) (2010) 1018–1021
5. Fano, R.: Transmission of information: a statistical theory of communications. M.I.T. Press (1961)
6. Jensen, B., Larsen, J., Jensen, K., Larsen, J., Hansen, L.: Estimating human predictability from mobile sensor data. In: IEEE International Workshop on Machine Learning for Signal Processing (MLSP) 2010, IEEE (2010) 196–201
7. Lin, M., Hsu, W., Lee, Z.: Predictability of individuals' mobility with high-resolution positioning data. In: Proc. of Ubicomp'12, ACM (2012) 381–390
8. Yan, X.Y., Han, X.P., Wang, B.H., Zhou, T.: Diversity of individual mobility patterns and emergence of aggregated scaling laws. Scientific reports **3** (2013)
9. Cho, E., Myers, S., Leskovec, J.: Friendship and mobility: user movement in location-based social networks. In: Proc. of KDD'11. (2011) 1082–1090
10. Kontoyiannis, I., Algoet, P., Suhov, Y., Wyner, A.: Nonparametric entropy estimation for stationary processes and random fields, with applications to english text. IEEE Transactions on Information Theory **44**(3) (1998) 1319–1327