Assessing spatiotemporal predictability of LBSN: A case study of three Foursquare datasets

Abstract. Location-based social networks (LBSN) provide new possibilities for researchers to gain knowledge about human spatiotemporal behavior, and to make predictions about how people might behave through space and time in the future. An important requirement of successfully utilizing LBSN in these regards is a thorough understanding of the respective datasets, including their inherent potential as well as their limitations. Specifically, when it comes to predictions, we must know what we can actually expect from the data, and how we could maximize their usefulness. Yet, this knowledge is still largely lacking from the literature. Hence, this work explores one particular aspect which is the theoretical predictability of LBSN datasets. The uncovered predictability is represented with an interval. The lower bound of the interval corresponds to the amount of regular behaviors that can easily be anticipated, and represents the correct predication rate that any algorithm should be able to achieve. The upper bound corresponds to the amount of information that is contained in the dataset, and represents the maximum correct prediction rate that cannot be exceeded by any algorithms. Three Foursquare datasets from three American cities are studied as an example. It is found that, within our investigated datasets, the lower bound of predictability of the human spatiotemporal behavior is 27%, and the upper bound is 92%. Hence, the inherent potentials of the dataset for predicting human spatiotemporal behavior are clarified, and the revealed interval allows a realistic assessment of the quality of predictions and thus of associated algorithms. Additionally, in order to provide further insight into the practical use of the dataset, the relationship between the predictability and the check-in frequencies are investigated from three different perspectives. It was found that the individual perspective provides no significant correlations between the predictability and the check-in frequency. In contrast, the same two quantities are found to be negatively correlated from temporal and spatial perspectives. Our study further indicates that the heavily frequented contexts and some extraordinary geographic venue types such as airports could be good starting points for effective improvements of prediction algorithms. In general, this research provides novel knowledge regarding the nature of the LBSN dataset and practical insights for a more reasonable utilization of the dataset.

Keywords: predictability; spatiotemporal behavior; context; location-based social networks; Foursquare; citizen sensing;

1 Introduction

Gaining knowledge on human spatiotemporal behavior has ever been a perennial research topic. Nowadays, due to its abundant potential applications, it is especially valuable even from a more practical point of view. Exemplary application areas include rather different fields such as analyses of public transit flows (Steiger et al. 2014), pervasive advertising (Ghafourian and Karimi 2011, Spiegler et al. 2011), route planning (Gu et al. 2014, Zhu et al. 2014), location recommendation (Gavalas and Kenteris 2011, Majid et al. 2013) or disaster management (de Albuquerque et al. 2015).

In order to carry out such research, data proxies that are capturing human behavior are needed. From a historical point of view, one of the earliest data proxies capturing human behavior was bank notes. Given that these are used day-to-day and in ordinary situations, there is some tradition of analyzing bank note dispersal (e.g., Brockmann et al. 2006). Nevertheless, one of the most widely adopted data proxy to study the human spatiotemporal behavior during the past decades should be the mobile phone data. Thanks to the GPS technology and the popularization of mobile phones, it is very convenient to recover individual trajectories from the mobile phone usage data in a rather large scale. Consequently, plenty of studies have been conducted to model and predict human spatiotemporal behaviors based on this type of data proxy (González *et al.* 2008, Barabási 2011, Giannotti *et al.* 2011, Parent *et al.* 2013, Do *et al.* 2015).

The recent years, however, have witnessed a dramatically changed lifestyle of modern society. Along with the ubiquitous access to the Internet and the popularity of various kinds of location-based social networks (LBSN), people are increasingly willing to report their personal experiences on the social networks from their immediate vicinity in all kinds of situations. These reports can be exceptional happenings, but mostly are very ordinary everyday situations. Hence, this alleged trivial information allows researchers to observe the human behavior up to a certain level of detail. Popular examples of LBSNs include, for example, the microblogging service Twitter, the personal communication hub Facebook and the check-in service Foursquare. The technological and the cultural change have conjointly provided an ample amount of detailed insights into the users' everyday life that was never feasible before.

Given the pervasive nature of such services, the corresponding data sources are nowadays becoming a popular proxy for reflecting human behavior (see Lee and Sumiya 2010, Preoţiuc-Pietro and Cohn 2013, Liu *et al.* 2014).

From the perspectives of citizen sensing (Goodchild 2007, Sheth 2009), each person in these LBSNs can be regarded a volunteered sensor to report any aspects of a citizen's daily life. However, unlike traditional sensors which are often carefully designed and calibrated in order to deliver accurate and homogeneous measurements, the volunteered social sensors act autonomous and subjective. Therefore, the information they deliver is typically fuzzy, uncertain and incomplete (Sengstock *et al.* 2013). This clearly constitutes a major obstacle towards utilizing these datasets properly. In fields involving more traditional datasets, researchers often have access to technical specifications that explain the quality and granularity of the data at hand. Consequently, assessments of the quality of any achieved results are possible, and are an asset in these cases. With LBSN and other social media feeds, in contrast, this is typically hampered by the challenges mentioned above. Nonetheless, proving the validity of scientific results is just as important as the novelty of approaches and applications.

Beyond the numerous studies that employ social media datasets to identify events or city structures (Liu *et al.* 2011, Sun *et al.* 2016), make recommendations of routes or venues (Noulas *et al.* 2012, Kurashima *et al.* 2013), predict human behavior or interest (Noulas and Scellato 2012, Li *et al.* 2016), etc., one can indeed find some studies that investigate the dataset itself. For example, Cramer *et al.* (2011) investigated the Foursquare dataset by analyzing 20 in-depth interviews with Foursquare users, and discussed the performance aspects as well as some norms and conflicts in the dataset. By correlating Twitter data with the UK census data, Steiger *et al.* (2015) explored the semantic associations between tweets and their respective spatiotemporal whereabouts, and examined the potentials of Twitter being an indicator for people's whereabouts. With the help of face-to-face interviews and usage of external datasets, these studies have provided valuable insights into the utilization of social media data.

In this study, as the first goal, we intend to supplement these insights with a kind of intrinsic property of the dataset that indicates the potentials and limits for making predictions about human's spatiotemporal behavior, e.g., *predictability*. In this paper, it is defined as the degree to which a correct prediction of user's spatiotemporal behavior can be made at best and at worst based on some given dataset (see Section 2 for a more thorough definition of predictability). It is quantitatively represented as an interval, where the lower bound corresponds to the regular behaviors and the upper bound corresponds to the information amount as is contained in the corresponding dataset. These bounds thus describe the *prospects of the dataset* with respect to predictions irrespective of any specific algorithm used. The second goal of our study is to provide some useful guidance for practical scenarios. Therefore, we also investigate the relationships between predictability and the check-in frequencies from three different perspectives: individual, temporal and spatial. The revealed relationships can hint algorithm designers to those parts of their algorithms that are worth being tuned given a specific dataset at hand. All analyses in this paper are undertaken on three exemplary Foursquare datasets which originate from the three most populous US cities: New York City, Los Angeles and Chicago. The obtained empirical results and suggestions can act as important inputs towards drawing better predictions from such datasets. They allow evaluating prediction algorithms more realistically, and they further help in increasing their quality.

We start our paper by introducing the formal definition of predictability as well as our assessment approach in Section 2. Afterwards, the employed datasets as well as the preprocessing steps are described in Section 3, and the assessment results are presented in the same section. Section 4 unravels the relationships between predictability and the check-in frequencies from different perspectives. Some recommendations on how to efficiently improve prediction algorithms are given accordingly. Section 5 concludes the paper by discussing the results.

2 Predictability: definition and assessment

As stated earlier, predictability is supposed to reflect the intrinsic interval of prediction accuracy based on the data itself. Its definition and assessment should therefore not be restricted by any specific predicting algorithms, but should rather be understood as the "potential of a dataset". In this chapter, we lay out our formal definition and assessment approaches of predictability.

In 2010, Song *et al.* studied human mobility from the trajectories from mobile phone usage data, and assessed the predictability of their data. Their research also forms the basis of our study. Therefore, the relevant part of their work will also be introduced in the related subsections.

2.1 Formal definition of predictability

In their work, Song *et al*. laid out a formal definition of predictability of mobile phone data. The starting point is a temporally ordered personal trajectory history of length $n-1$ that is represented as $h_{n-1}=[X_{n-1},X_{n-2},\cdots,X_1]$, where X_i denotes the user's location at time i. They define $\pi(h_{n-1})$ as the probability that a user is at his/her most likely location given his/her trajectory history h_{n-1}, that is,

$$\pi(h_{n-1})=\sup_x\{\Pr[X_n=x\,|\,h_{n-1}]\}. \quad (1)$$

The authors then explain that $\pi(h_{n-1})$ contains the full predictive power that is present in the data. Afterwards, they sum over all possible trajectories of length $n-1$, and define the predictability $\Pi(n)$ for the trajectories of history length $n-1$ as:

$$\Pi(n)=\sum_{h_{n-1}} p(h_{n-1})\cdot\pi(h_{n-1}), \quad (2)$$

where $p(h_{n-1})$ denotes the probability of observing a particular trajectory history h_{n-1}.

Equation (2) is a function of the trajectory length, while different users possess trajectories of different lengths. In order to capture the overall predictability within the entire dataset, the authors define the overall predictability by taking the limit of Equation (2), which leads to:

$$\Pi=\lim_{n\to\infty}\frac{1}{n}\sum_i^n\Pi(i) \quad (3)$$

Song's work unraveled the potentials of mobile phone data for studying human mobility, and provided an answer to their research question "to what degree is human behavior predictable". Their findings also hint on many follow-up studies of human mobility based on mobile phone datasets (see e.g., Salah *et al.* 2010, Calabrese *et al.* 2011). However, Song's approach cannot be directly applied to social media feeds because of the significant differences between the two kinds of datasets. On the one hand, mobile phone data comes up to a very high and more or less balanced coverage. That is, a broad range of people are using these services regularly, and the dataset is less prone to being dominated by single individuals. Contrarily, according to Li *et al.* (2015), more than 80% of the entire Foursquare check-ins are contributed by merely 1% of the entire user community. Therefore, it is, for a great deal of users, not feasible to recover the individual personal trajectories from such a "heavy-tail distributed" dataset. However, the personal trajectory is the most fundamental element in Song's work. On the other hand, in comparison to mobile phone datasets where locations are usually represented as simplified geometric points, social media datasets contain rich semantic and contextual information, and these semantics of locations or contextual conditions in turn have great influences on user behaviors. For example, McKenzie *et al.* (2015) proves that the temporal characteristics of human behavior differ significantly with the semantics of the place, or the place type.

Hence, in this work, we adapt Song's approach to the social media datasets. Instead of considering predictability from personal trajectory history, our adapted definition of Π is based on the contextual conditions. This follows from the fact that semantic venues enriched by contextual information such as location and time have been massively explored with check-in datasets (see McKenzie *et al.* 2013, Krueger *et al.* 2014). Put formally, the adapted version writes as

$$\Pi=\sum_{c_i\in\mathrm{C}} p(c_i)\pi_i, \quad (4)$$

where each element $c_i\in\mathrm{C}$ is either an atomic type (e.g., space or time) or a composition of various types of contextual information (e.g., tuples consisting of both temporal and spatial information). In addition, π_i in Equation (4)**Error! Reference source not found.** is the probability that a user checks in at the most likely venue given the observed contextual condition c_i. It is formally defined as:

$$\pi_i=\sup_v\{\Pr[V_i=v\,|\,c_i]\}, \quad (5)$$

where V_i is the user's check-in venue given the contextual condition c_i.

Hence, π_i represents the theoretical limit of the probability of making correct predictions given the realization of a certain type of contextual condition c_i, and Π represents the overall probability of the dataset across all possible kinds of contextual conditions. Since the definition does not concern any details of specific prediction algorithms, Π can be regarded as a measure of the inherent predictability contained in the dataset itself, rather than an evaluation of some algorithm's predictive power. In the next two subsections, we will explore the bounds of Π ($\Pi^{\min}\le\Pi\le\Pi^{\max}$).

2.2 The upper bound

In Song's work, the upper bound is determined based on the Shannon's entropy (Shannon 1948) and Fano's inequality (Fano 1961), which are written respectively in Equation (6) and (7):

Shannon's entropy:

$$H(X) = \sum_{x \in \mathrm{X}} -p(x) \log p(x), \tag{6}$$

where $p(x)$ is the probability mass function of the random variable X.

Fano's inequality:

$$H(X \mid Y) \le H(e) + p(e) \log(N-1) \tag{7}$$

where $p(e) = \Pr(X \neq \hat{x}_a)$ is the probability of making erroneous predictions, $H(e)$ is the corresponding entropy following Equation (6), and N is the number of all possible predictions. In addition, $H(X \mid Y)$ is the conditional entropy of a random variable X given the knowledge of another random variable Y:

$$H(X \mid Y) = \sum_{x \in \mathrm{X}} \sum_{y \in \mathrm{Y}} -p(x \mid y) \log p(x \mid y)\, p(y) \tag{8}$$

Based on Equations (6) and (7), Song *et al*. proved that $\Pi^{\max}$ can be determined with the following Equation (8):

$$H(X \mid Y) = -\Pi^{\max} \log \Pi^{\max} - (1-\Pi^{\max}) \log(1-\Pi^{\max}) + (1-\Pi^{\max}) \log(N-1) \tag{9}$$

With the term predictability redefined in our context, a direct application of Equation (8) and (9) into social media datasets would yield problems. Due to the heavy-tailed distribution, the sample size would be too small for plausible estimation of conditional probability $p(x \mid y)$ and thereby $H(X \mid Y)$. Therefore, the estimation of $H(X \mid Y)$ must be corrected to cope with the small sample issues. Here we apply the Miller-Madow bias correction (Miller 1955) (the $1/2N$ summand in Equation (10)), which is the most classic correction approach to solve the small sample issues. This additional term positively corrects the otherwise appearing underestimation of the entropy. Hence, we get the corrected estimation of $H(X \mid Y)$ as Equation (10), where the random variable X is represented with the unknown venues V, and the conditional random variable Y is represented with the known contextual conditions C:

$$\begin{aligned} H(V \mid C) &= \sum_{c_i \in \mathbf{C}} p(c_i) \sum_{v_j \in \mathbf{V}} \left(-p(v_j \mid c_i) \log p(v_j \mid c_i) + \frac{1}{2N} \right) \\ &= \sum_{c_i \in \mathbf{C}} \frac{\lambda_{c_i}}{\lambda} \sum_{v_j \in \mathbf{V}} \left(-\frac{\lambda_{v_j, c_i}}{\lambda_{c_i}} \log \frac{\lambda_{v_j, c_i}}{\lambda_{c_i}} + \frac{1}{2\lambda_{c_i}} \right) \end{aligned}. \tag{10}$$

Here λ_m is the number of check-ins under some given condition m. Together with Equations (9) and (10), the upper bound of predictability of social media datasets can then be determined.

2.3 The lower bound

As stated by Song *et al*., "not only that a certain amount of randomness governs their future whereabouts, but also that there is some regularity in their movement that can be exploited for predictive purposes." In their research, Song *et al*. firstly defines regularity at the n-th step $R(n)$ as the expected probability that the user is in his/her most likely position given the observed history h_{n-1}. That is, regularity is measured as $R(n) = \Pr(X = x_{ML} \mid h_{n-1})$, and the overall regularity R (i.e., regardless of n) is achieved by taking the limit, i.e., $R = \lim_{n \to \infty} \frac{1}{n} \sum_{i}^{n} R(i)$. The authors further proved that $R \le \Pi$, and therefore the value of R represents the lower bound of predictability.

The regularities of human spatiotemporal behavior can be well illustrated by the regular day-to-day routine of people. For example, if a user checks-in at Starbucks every morning between 8 and 9am, this information will be reflected in the data by principle, regardless of the prediction performance. Therefore, even the simplest prediction algorithm will be able to exploit this information. Thus, it is also intuitively comprehensible to capture the lower bound of predictability with regularity.

Similarly, we define regularity as the expected probability of finding a user in his/her most likely visited venue given a specific contextual condition, which can be written as:

$$\Pi^{\min} = R = \sum_{c_i \in C} p(c_i) \Pr(X = x_{ML} \mid c_i), \tag{11}$$

where x_{ML} is the most likely visited venue conditioned on the contextual condition c_i. Thus, if a user does possess regular behavior like described above, this value will be relatively high. Elsewise, if a user behaves rather random, the value will be low, meaning that algorithms would need to be more sophisticated for achieving useful results.

The determination of the conditional probability $\Pr\left(X = \hat{x}_{ML}^{c_i} \mid c_i\right)$ is challenging again. The traditional estimation of p by using the available information is relying on large-sample theory, and thus convergence of the estimated parameter $\hat{p}$ against the true one. However, as we are dealing with small sample sizes, we must incorporate the estimation error / uncertainty into our considerations. One of the very commonly applied algorithms to achieve this is the so-called Wilson score interval. Further, as we are working with multinomial data, we correct the probability estimation with the corresponding multinomial version this technique (Wilson 1927), and estimate the conditional probability with the following equation:

$$\hat{p} = \frac{n}{n+z^2} \cdot \frac{\|x_{ML}\|}{n} + \frac{z^2}{n+z^2} \cdot prior, \tag{12}$$

where z is the $1-\frac{1}{2}\alpha$ percentile of a standard normal distribution, α is the confidence level, n is the sample size conditioned on the context c_i, and $prior$ describes the a priori knowledge of the targeted venue $\hat{x}_{ML}^{c_i}$ that is extracted from all users under that context. Taking the 95% confidence interval, i.e., $\alpha = 0.05$, and thus $z = 1.96$ (since the Wilson score is approximating the bounds by a standard normal distribution), the lower limit of predictability can be determined from Equations (11) and (12).

3 A case study of Foursquare

In this section, we assess the bounds of predictability as outlined in the previous section with three Foursquare datasets from three major American cities (see Section 3.1). Due to the pronounced data sparseness, some preprocessing steps are necessary to ensure the statistical validity of our achieved results (Section 3.2). The estimated results of predictability for the examined datasets following the approach explained in Section 2 are then presented in Section 3.3.

3.1 Investigated datasets

Our experiment is carried out based on the Foursquare datasets. Foursquare offers the great advantage of following a well-defined hierarchy with respect to their venue descriptions. This makes Foursquare check-ins highly valuable for analyzing human spatiotemporal behavior, since it allows accurate characterizations of the visited venues. This advantage is even more valuable when considering that the hierarchy is not an authoritatively imposed one, but has been agreed on by the users themselves and became only standardized globally later on. Thus, the particular categories reflect (to a certain extent) what is relevant to the users instead of some subjective organization.

Our data originates from the three most populous American cities: Chicago, Los Angeles and New York City. Each check-in record from the dataset includes information on time, location and user as well as profiles of the venues attached to the respective check-ins (e.g., name, category and subcategory). In total we have collected 183,837 (Los Angeles), 138,211 (Chicago) and 579,786 (New York City) check-in records during a period of five months from February 1st to June 30th 2014.

3.2 Preprocessing: aggregation of the sparse data

Similar to many other LBSN datasets, the Foursquare check-in dataset is large but sparse. That sparseness is caused by the fact that a relatively small percentage of highly active users accounts for a large portion of the overall data. The cumulative distribution function (CDF) in Figure 1 shows that more than 90% of all users create fewer than 30 records during the 150 observed days. Specifically, we found that the distribution of the individual numbers of check-ins over all users is well approximated by a truncated power law:

$$\Pr(x) = (x + x_0)^{-\beta} \exp^{-x/\kappa}, \tag{13}$$

which is parameterized as follows: $\beta^{Ch}=1.7, \beta^{LA}=1.71, \beta^{NYC}=1.59$; $x_0^{Ch}=0.98, x_0^{LA}=0.82, x_0^{NYC}=1.4$, and the cut-off value $\kappa^{Ch}=291, \kappa^{LA}=103, \kappa^{NYC}=157$.

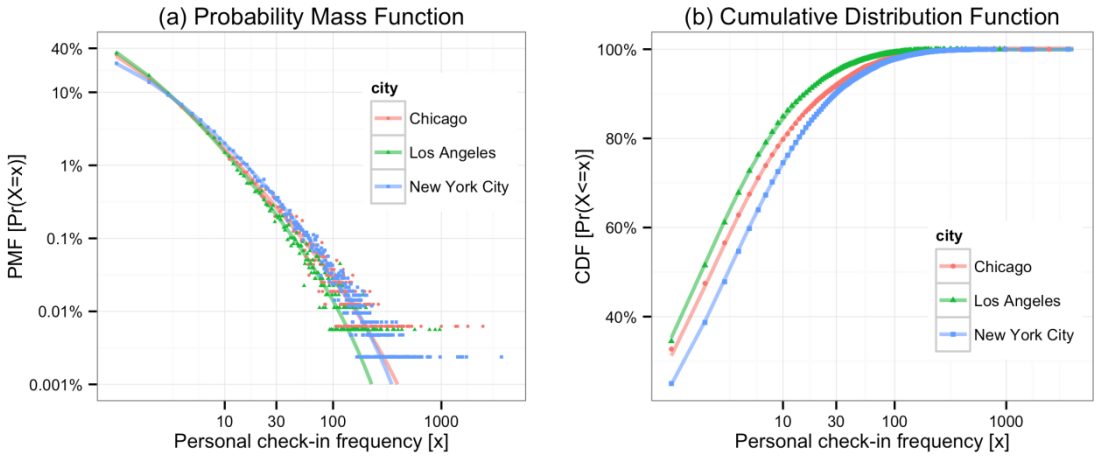

Figure 1. A truncated power law distribution of users' personal check-in frequency for all three cities. a) probability mass function, b) cumulative distribution function.

Based on such sparse data, it is problematic to draw valid statistical conclusions. In order to deal this issue, it is pretty much common sense to filter out inactive users based on some criterion (Rattenbury *et al.* 2007, Quercia and Lathia 2010). This approach is indeed simple and effective; but the research is then limited to a rather small portion of the overall users which dominate any results. Further, the dataset is not leveraged to its full potential. Another way to deal with this problem is collaborative filtering, i.e. to learn knowledge of one user from other similar users (Pham *et al.* 2011, Ye *et al.* 2011, Liu *et al.* 2013). We are convinced that the data from inactive users also contributes to a thorough understanding of users' spatiotemporal behavior. That is, in order to describe the whole dataset, we must include all constituting parts, which include these users too. Therefore, we will aggregate the data by considering user similarity.

Suppose each user u_i is characterized by a visit vector $\mathbf{U}_i=\left[n_i^1, n_i^2, \cdots, n_i^J\right]^T$, where each element n_i^j indicates the number of visits of venue j by user u_i. The user similarity is then measured by the cosine distance between their visit vectors. Such design follows one of the most frequently used approaches to measure user similarity (see Woerndl *et al.* 2009, Cho *et al.* 2011, Noulas *et al.* 2012). Thereafter, k-means is adopted to cluster the users based on their similarity.

The choice of k is the crux to k-means clustering. In order to find a suitable k, we carry out some preliminary experiments. On the one hand, we observe the degree of statistical reliability after clustering. This is measured with the proportion of clusters that contain enough records for statistical purposes, e.g., 30. On the other hand, we observe the preservation of heterogeneity. This is measured by the ratio of *Between Sum of Squares* (BSS) and *Total Sum of Squares* (TSS). As is known from the analysis of variance (ANOVA), the total sum of squares (TSS) is the sum of the so-called "within-samples" sum of squares (WSS) and "between-samples" sum of squares (BSS), and WSS is a representation of the intra-cluster heterogeneity. After clustering, the records in the same cluster will be treated equally, thus the WSS will be neglected. Hence, a higher ratio of BSS and TSS indicates a smaller WSS to be neglected; in other words, more heterogeneity can be preserved.

Figure 2(a) shows some exemplary outcomes of this pre-analysis derived from Chicago data. As the number of clusters (k) decreases, each cluster will include a larger number of check-ins, which generally leads to more reliable statistical results. At the same time the ratio of BSS and TSS is decreasing, which means more intra-cluster heterogeneity has been sacrificed. Hence, the actual decision of k involves a trade-off between the preservation of heterogeneity and the statistical reliability.

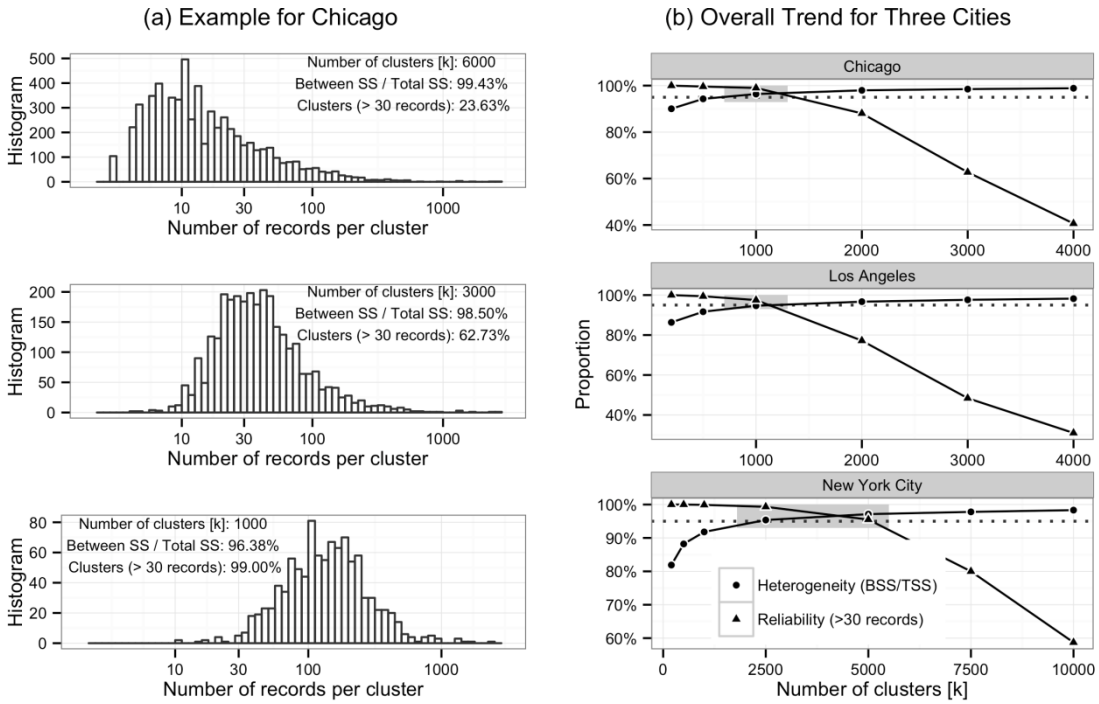

Figure 2. Identification of the number of clusters (k) based on two criteria: the degree of heterogeneity preservation (ratio of BSS and TSS) and the degree of statistical reliability (percentage of clusters containing more than 30 records). A more detailed variation of the two criteria with k is given in (a) for Chicago and the overall trend of variation is given in (b) for all three cities.

Figure 2(b) shows how the two criteria, i.e., the degree of heterogeneity preservation and the degree of statistical reliability, proceed along with k across all the three cities. Based on the available trends from Figure 2(b), k is set as 1000, 1000 and 2500 respectively for the three cities, so that over 95% of the clusters have more than 30 check-ins, while over 95% of the heterogeneity in the data remains preserved. This leads to clusters that subsume users that are showing similar check-in activity. In the following we consider each of those clusters as if they were one *individual*. This guarantees to have reasonable check-in numbers and allows more efficient analyses.

3.3 The assessment result of predictability

In this section, we present the assessment result of predictability. The predictability is assessed using the approaches described in Section 2 and conditioned on the spatial and temporal contexts. More specifically, the spatial condition will be represented by the zip code region of the check-in records, while the temporal condition will be represented by the hour slots of the check-ins. We consider the spatiotemporal information because this information has been heavily exploited in the existing studies of LBSN data, and its influences on human behaviors have been widely acknowledged.

The assessment results of the predictability of the three pre-processed Foursquare datasets across the three study sites are presented in Figure 3. The figure shows that the upper bounds of predictability conditioned on the spatiotemporal contexts are around 92% for all the three cities, while the lower bounds are around 27%. This indicates that about 27% of the spatiotemporal behaviors are rather regular in space and time and can easily be anticipated, while about 8% of them are totally random and cannot be predicted in theory

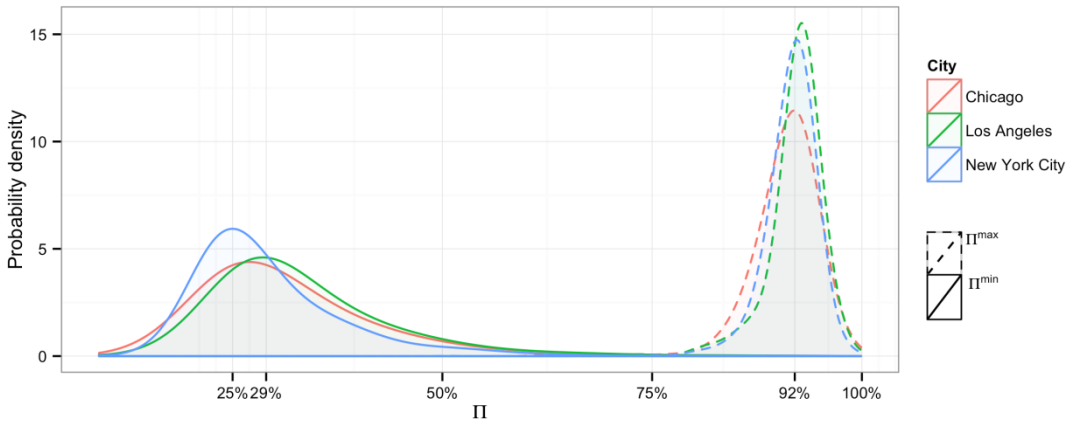

Figure 3. Probability density functions of the upper and lower bounds of the predictability

Put into simple words, these findings can be interpreted as follows: by considering spatiotemporal contextual information any prediction algorithm will at least make around 27% correct predictions on average, because 27% of the dynamics that are exhibited by the datasets are governed by very regular activities. Beyond the initial 27%, the users' spatiotemporal behavior can still be predicted with optimized algorithms, because the dataset contains further information than the obvious regularities. The algorithm can be tuned up to a maximum value of 92%. The remaining 8% are "unpredictable," because they are governed by randomness such as some emergent behavior. Hence, for example, if for any prediction algorithm a prediction accuracy of 30% is being stated, it means that besides the inherited 27% regularity in user's spatiotemporal behavior, the optimization of an algorithm is actually responsible for the extra 3% gain in accuracy. Furthermore, we would also know that there is still much room left for this algorithm to improve its accuracy.

A follow-up question that occurs naturally will then be: How can one achieve these improvements? We assume that the precision of prediction is not equal across different kinds of contextual conditions. Instead, the overall precision is to some degree a result of neutralization (some conditions contribute positively; some others might lower the achieved precision). If the variation patterns of predictability underlying different conditions could be revealed, one might gain better knowledge on how to improve some algorithm. This information will be revealed in the following Section 4.

4 Practical implications

Our study does not only focus on the mere theoretical numbers of the bounds of predictability. We also consider the practical implications of these values. In this section we investigate the relationships between predictability and the check-in frequencies from three perspectives (individual, time and space),, hoping to provide some further insights for future studies on prediction algorithms. Three predictability-related quantities are involved in this investigation: Π^{max}, Π^{min} and Π^{delta}. Here Π^{max} and Π^{min} are the upper and lower bounds as determined in the previous Section 3, and thus the difference $\Pi^{max} - \Pi^{min}$ gives the theoretical room for improvement on an absolute scale. Since we care more about the leverage effects of the partial improvements for the overall precision, i.e. the improvement efficiency, a related relative quantity Π^{delta} is defined as $\left(\Pi^{max} - \Pi^{min}\right) / \Pi^{min}$ to capture the theoretical efficiency of improvement.

4.1 Individual perspective

In this section, we attempt to clarify the relationship between predictability and check-in frequency from the perspective of individuals. It appears quite reasonable to assume some kind of relationship between predictability and check-in frequency: either by assuming a positive relationship (more information = better informed predictions) or a negative one (more information = more fuzzy predictions). However, Figure 4 reveals that these quantities are unrelated. This finding holds for all three investigated datasets. Note that each dot in Figure 4 represents a cluster of users that are aggregated together in the preprocessing step because of their similarity.

Figure 4. Variations of bounds of predictability (Π^{max} , Π^{min}) and improvement efficiency (Π^{delta}) with check-in frequency from an individual perspective.

One hypothesis to explain the dispersed nature of the scatterplots in Figure 4 is that users might show some similar check-in patterns, regardless of how frequently they are actually using the service. If this were true, it would be quite reasonable to leverage the "collective wisdom" under certain spatiotemporal contexts to predict the behavior of the inactive users or even new users. Another hypothesis is that the more active users might provide more information from which to be learned, which is good for the purpose of predicting. However, their activities tend to be more dynamic, which increases the challenge in predicting. Thus the final dispersed distribution could result from both perspectives. Either way, Figure 4 shows that the common practice of filtering inactive users (Rattenbury *et al.* 2007, Quercia and Lathia 2010) is not quite efficient for improving the overall performance of a prediction algorithm.

4.2 Temporal perspective

Figure 5 depicts the relationship between the three predictability-related quantities and the check-in frequency from the perspective of time. A negative correlation can be found for both bounds of predictability (Π^{max} , Π^{min}), while a positive one is found for theoretical improvement efficiency (Π^{delta}).

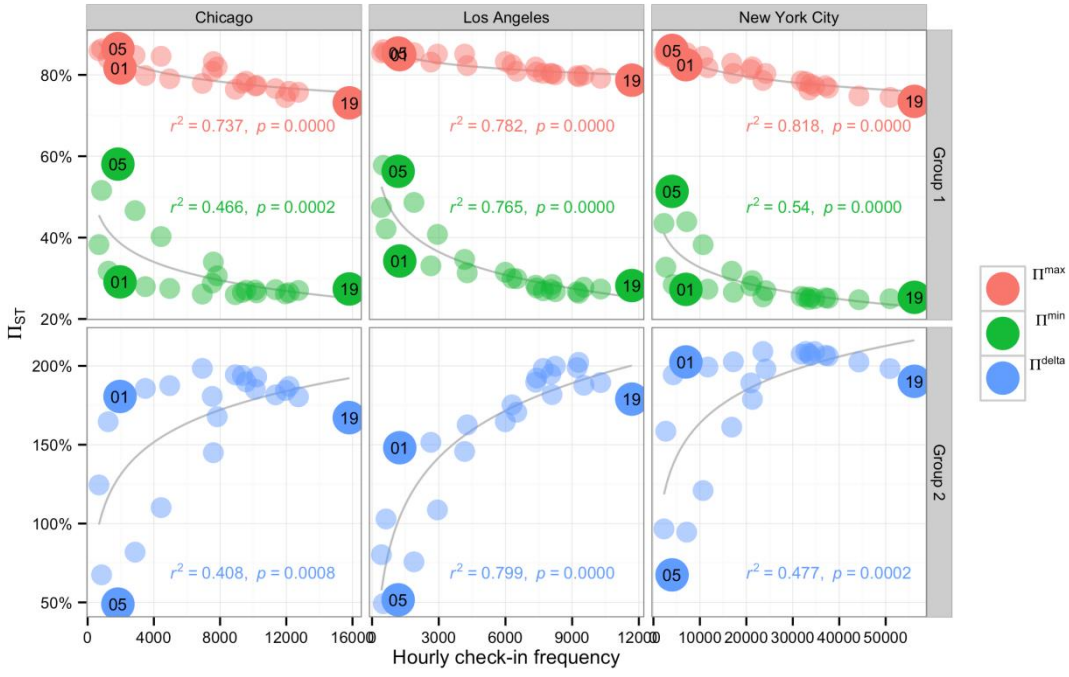

Figure 5. Temporal variation patterns of predictability in forms of the upper bounds (upper three subplots), lower bounds (three subplots in the middle) and improvement efficiency (lower three subplots). The labelled dots relate to the time slots explained within the subsequent Figure 6.

The negative correlations in the upper three subplots suggest that busier hourly slots provide poorer initial prediction accuracy. In other words: these slots still provide plenty of room for refining the prediction accuracy. Additionally, the positive correlations shown in the lower three subplots suggest that busier hourly slots also provide stronger leverage effects with respect to improving algorithms. Both of these aspects indicate that focusing on the busy time slots can help to diagnose and improve a prediction algorithm both effectively and efficiently.

In order to further unravel the reason why busy hours correspond to poor predictability, Figure 6 explores the semantical composition of three typical time slots (a lower outlier "01", an upper outlier "05" and a tail point "19" in Figure 5). The temporal semantics are represented by the probability mass function over the check-in categories within the respective time slots.

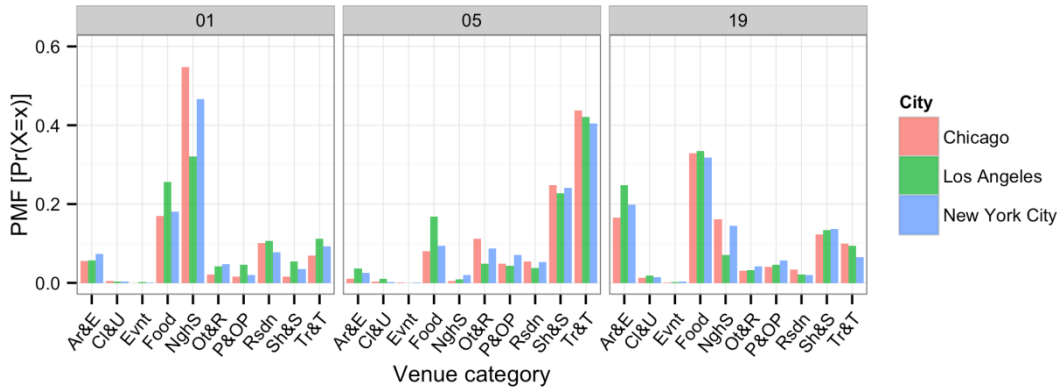

Figure 6. Temporal semantics: probability mass function (PMF) of the categories of check-in venues in three typical hours. Note that these hours (numbers in the head of the figure) do actually relate to real hours of the day.

According to Figure 6, the reason for higher predictability in less busy hours could be related to simpler temporal semantics. During 1 am and 5 am, the users rarely check in at venues of categories other than *Nightlife Spot* and *Travel and Transport*. Correspondingly, these time slots tend to have a high initial predictability, while the room for improvements is quite narrow. In other words: if the users just check in at a limited number of venue types it is quite likely to achieve precise predictions. In contrast, during the busy hours, users tend to pursue much more diverse activities and the temporal semantics are much more complex for predictions, see the temporal semantics for 19 pm in Figure 6.

Therefore, the indication here is that more effort should be invested in further enriching the busy hours with more detailed contextual information. In contrast, the predictions for hours showing low activity rates might be rather accurate by just simply considering temporal contextual information. Improving the prediction accuracy for these hourly slots is not worth the effort. Fortunately, the busy hours come with a much richer wealth of data to learn from; therefore it is also practical to shift the focus to these busy time slots only.

4.3 Spatial perspective

In analogy to the temporal characteristics we also investigate the relationship between spatial check-in frequency and the three predictability-related quantities (Figure 7). Here, we are using zip code areas as our spatial units.

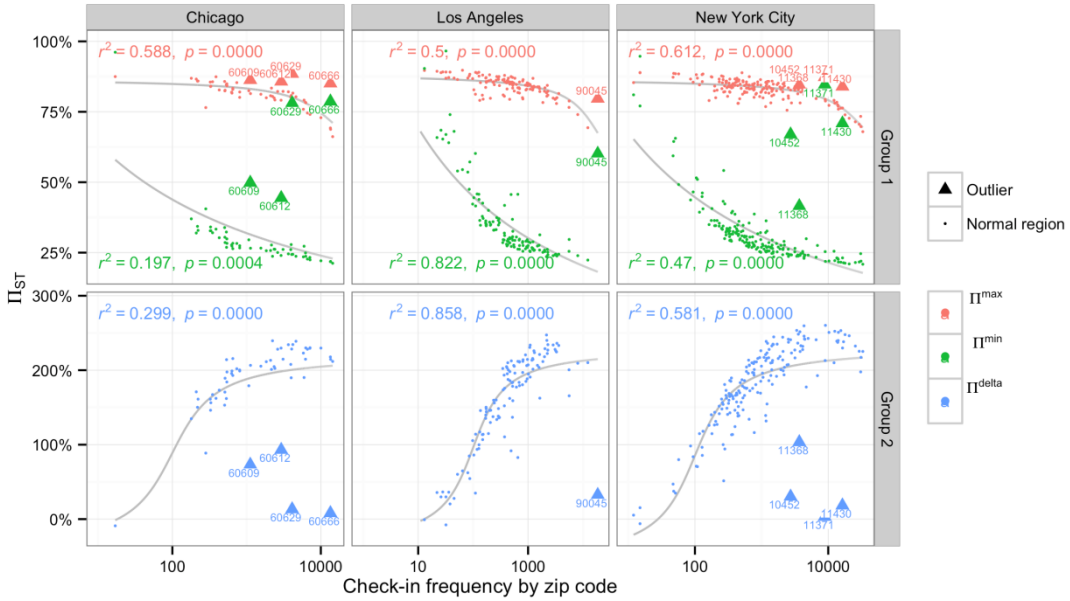

Figure 7. Spatial correlation of check-in frequency with the upper and lower bounds (upper three subplots) and with the improvement efficiency (lower three subplots) for three investigated datasets respectively. Each dot in the figure corresponds to one zip code area.

The gray lines in Figure 7 capture the overall functional relationship between the check-in frequency and predictability. In general, the heavily frequented spatial regions are associated with lower predictability and better improvement efficiency. Hence, Figure 7 from the spatial perspective reveals quite similar patterns as that of Figure 5 from the temporal perspective. Therefore, we assume similar underlying reasons to be effective: the heavily frequented spatial regions also tend to contain complicated semantics because of the complex functions they typically provide. Hence, the initial prediction accuracy is low while the efficiency for improvement is quite high.

However, compared with the temporal correlations in Figure 5, Figure 7 distinguishes itself by showing several noticeable outliers which lie far from the central gray line. These outliers are represented by triangular dots. In these outlier zip code areas, users are frequently checking in, yet the predictability is still quite high. Thus, it would be interesting to further investigate these outliers and find out why these regions do not follow the general trend. Hence, Figure 8 inspects the spatial semantics of the outliers with respect to their categorical distributions.

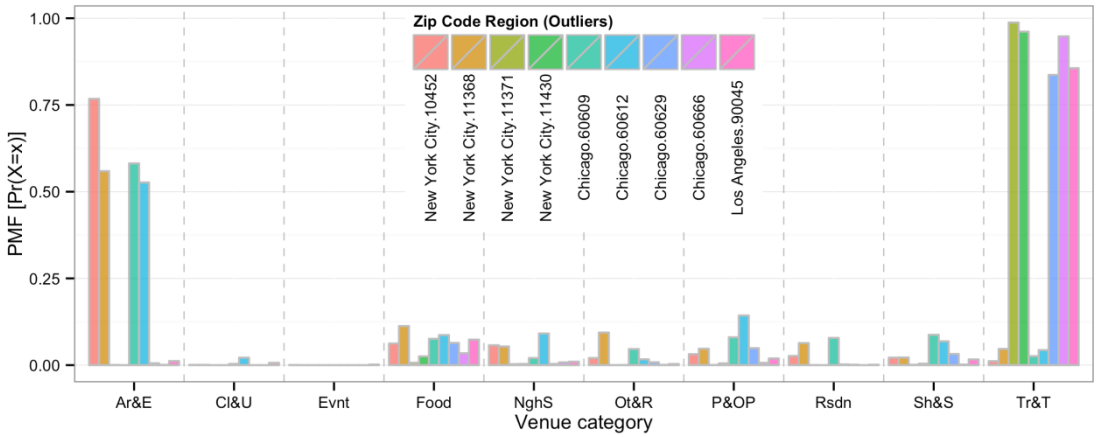

Figure 8. Spatial semantics: probability mass function of the categories of check-in venues in the outlier regions as highlighted in Figure 7.

Figure 8 shows that, regardless of the outlier category, these outlier regions are dominated by just one single semantic category each (e.g., *Travel and Transport*, *Arts and Entertainment*). Hence, these

outliers actually confirm our assumptions that were raised earlier: simpler semantics result in higher initial prediction accuracy while providing lower improvement efficiency.

Table 1 An overview of the outlier regions.

	City	Postal Region	Geographic Phenomena	Dominant Venue
1	Chicago	60609	Baseball park	US Cellular Field
2	Chicago	60612	Sports arena	United Center
3	Chicago	60629	Airport	Chicago Midway International Airport
4	Chicago	60666	Airport	O'Hare International Airport
5	Los Angeles	90045	Airport	Los Angeles International Airport
6	New York City	10452	Baseball park	Yankee Stadium
7	New York City	11368	Baseball park	Citi Field
8	New York City	11371	Airport	LaGuardia Airport
9	New York City	11430	Airport	John F. Kennedy International Airport

When the outlier instances are all mapped back into geographic space, it is found that each outlier region contains either a large airport or a large sports stadium (see Table 1). The special types of geographic venues attract huge amount of users meanwhile reducing the diversity of human behavior in their corresponding localities. Therefore, together with Figure 8, Table 1 shows how tremendously certain venue types can impact human's spatiotemporal behavior.

The dominance of some venue types can be regarded as a general phenomenon. For instance, we can expect international airports to attract an enormous number of check-ins in most cities, even well beyond our study sites. In contrast, the dominance of some other types of venues might be explicitly local effects; and can therefore not be generalized to a universal viewpoint as easily. Baseball, for example, is one of the most popular sports in the United States. Therefore, baseball parks strongly impact human's spatiotemporal behavior across this particular country. Compared to New York City, however, the users in Chicago do also show a great passion on basketball. Therefore the *United Center*, which is the home to the *Chicago Bulls* of the National Basketball Association (NBA), dominates the spatiotemporal behavior in the corresponding zip code area. Anyway, it seems to be a valid statement that mass events like sports games should attract a large number of check-ins across the world.

5 Discussion and Conclusions

Researchers working with LBSN datasets are often confronted by themselves or others with doubts regarding the quality or the potential of their datasets. It is reasonable to be skeptical, indeed. Therefore, in this article, we investigate one aspect which is governing parts of the quality of LBSN datasets: their inherent predictability. Knowledge on the predictability can help researchers gaining a deeper understanding of their working datasets. Further, a more thorough anticipation of prediction algorithms as well as a more reasonable explanation of results drawn from their application can be made. Additionally, knowledge on the relationships between the check-in frequencies and the predictability can further hint on those parts of any algorithm that are worth being tuned in order to improve the quality of the drawn predictions. Our work therefore contributes such useful but yet missing knowledge about the predictability of LBSN datasets regarding these two aspects.

In the first part of our work we evaluate the bounds of the predictability by investigating three exemplary Foursquare datasets. This evaluation is based on the intrinsic amount of information contained in the respective datasets, as this is linked to the predictive power which a dataset provides. We found that the predictability of the spatiotemporal behavior attached with Foursquare datasets is bounded to an interval approximately between 27% and 92% for the three investigated datasets from three US cities (Chicago, Los Angeles and New York City). This finding is useful with respect to two aspects. On the one hand, these findings provide a way to assess existing prediction algorithms. An analyst can use the range mentioned above for comparing any actually achieved performance against it. Doing so allows a more realistic assessment of the quality of predictions and thus of associated algorithms. On the other hand, our findings also allow gaining a more comprehensive understanding of the investigated datasets. That is, we learn something about the informative value a dataset can provide. This includes information well beyond pure predicting. For instance, the lower bound of 27% does in turn also unveil that a considerable amount of users must show a regular spatiotemporal behavior. This is promising because it shows that people seem to at least partially integrate these services into their daily routines.

In the second part of our work we reveal the relationship between the check-in frequencies and the predictability from the perspectives of the individual, time and space. From the individual perspective it is found that the predictability is unrelated to individual check-in frequencies. This indicates that all available users can (and should) be used for making predictions without the need to filter out the less

active ones. In contrast, from the temporal and spatial perspectives, predictability is found to be negatively correlated with the check-in frequency. In other words: The downside of the more heavily frequented time slots and spatial regions is that these come up with more complex temporal or spatial semantics. It is thus more difficult to achieve good prediction results from these by employing trivial approaches. Meanwhile, however, these heavily frequented time slots and spatial regions do also provide the advantage of offering strong leverage effects on the prediction accuracy. Therefore, focusing on these particular contexts might be a good starting point for more effective improvements of prediction algorithms. Simply put, these slots and regions provide a lot of room for improvement. However, in some situations one also encounters peculiar situations that contradict these general rules and might appear as outliers. Further investigations on these outliers have shown that some extraordinary geographic venue types, such as airports and sports stadiums, have strong influences on spatiotemporal behaviors and thus the predictability values. An ad hoc treatment to these geographic phenomena is required in order to improve the overall quality of prediction algorithms.

In summary, the practical implications of our work are twofold. First of all, we serve analysts and algorithm designers with information about the extent of the intrinsic predictability that derives from LBSN datasets. That is, we provide the minimum and maximum possibilities with respect to the prediction performance in the presence of the predictive power of some specific underlying dataset. We further hint on the most effective characteristics that analysts should focus on in order to tune the prediction performance of their algorithms.

This paper focusses on predictions based on the spatiotemporal contextual conditions. Despite the essential role that the spatiotemporal information is playing in the related research, throughout the literature one can also find many prediction algorithms with emphases of other kinds of contextual information. For instance, Cheng *et al.* (2013) attempted to make predictions based on previously visited venues. The predictability values determined in the presented paper cannot be directly applied to such algorithms which are adopting other kinds of contexts. However, the approach of determining predictability that we proposed is not constrained to certain types of contexts (see, for example, Equation (6)). Readers might still extend our approach to determine their contextualized version of predictability measure fitted to their respective needs. It would further be an interesting future research task to assess the bounds of predictability on the background of interplay between our considered context types and others like the venues considered by Cheng *et al*.

In addition, we investigated the variation patterns of predictability from three perspectives: individuals, time and space. In practice, readers might face more specialized situations and may be interested in tailoring our proposed approach to their specific needs. In such case, readers can perform analyses similar to the approach presented in this work from other perspectives, or introduce their own situation-adapted approach as long as the analysis could provide them useful insight. Nevertheless, our work contributes to a better understanding of predictability in a rather general sense (i.e. beyond highly specialized application scenarios). Thus, we believe that the chosen focus on the spatial and temporal contexts is important for a broad range of studies on spatiotemporal behavior.

Furthermore, in this work, we leveraged the most basic information as provided by LBSN datasets, which is the spatial and temporal information. However, the LBSN datasets are also accompanied by some other types of information in the forms of texts, photos, networks, demographics, etc. These kinds of additional information are much more heterogeneous than the spatial and temporal information, while their potentials are not yet fully utilized in existing studies. In the future we are very interested in integrating this information into our research frame, and gain a more thorough understanding of the dataset.

Finally, in order to put our results into a broader context, we should mention that the representativeness of the spatiotemporal behavior from LBSNs with respect to the underlying comprehensive real-world human spatiotemporal behavior still remains mostly unknown to us (Ruths and Pfeffer 2014). Therefore, the predictability values detected from the LBSN datasets might not be applicable to the complicated human behavior in reality, just as the predictions drawn from LBSNs. Nevertheless, since the behavior represented by the dataset undeniably is one part of (contemporary) the real human behavior, it will definitely inherit some interesting features (e.g., the truncated power law distribution we have found in user contribution) and can be regarded as kind of a "window" to better understand the multi-dimensional real-world human spatiotemporal behavior.

Reference

Barabási, A.-L., 2011. Human Dynamics: From Human Mobility to Predictability. *In*: *Machine Learning and Knowledge Discovery in Databases*. Springer Berlin Heidelberg, 3–3.

Calabrese, F., Smoreda, Z., Blondel, V.D., and Ratti, C., 2011. Interplay between telecommunications and face-to-face interactions: a study using mobile phone data. *PloS one*, 6 (7), e20814.

Cheng, C., Yang, H., Lyu, M.R., and King, I., 2013. Where You Like to Go Next : Successive Point-of-Interest Recommendation. *In*: *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*. AAAI Press, 2605–2611.

Cho, E., Myers, S.A., and Leskovec, J., 2011. Friendship and mobility: user movement in location-based social networks. *In*: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*. New York, New York, USA: ACM Press, 1082.

Cramer, H., Rost, M., and Holmquist, L.E., 2011. Performing a check-in. *In*: *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services - MobileHCI '11*. ACM, 57.

Do, T.M.T., Dousse, O., Miettinen, M., and Gatica-Perez, D., 2015. A probabilistic kernel method for human mobility prediction with smartphones. *Pervasive and Mobile Computing*, 20, 13–28.

Fano, R.M., 1961. *Transmission of Information: A Statistical Theory of Communication*. Cambridge, Massachusetts: M.I.T. Press.

Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S., and Trasarti, R., 2011. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The International Journal on Very Large Data Bases*, 20 (5), 695–719.

González, M.C., Hidalgo, C.A., and Barabási, A.-L., 2008. Understanding individual human mobility patterns. *Nature*, 453 (7196), 779–82.

Goodchild, M.F., 2007. Citizens as sensors: web 2.0 and the volunteering of geographic information. *GeoFocus*, 7, 8–10.

Krueger, R., Thom, D., and Ertl, T., 2014. Visual Analysis of Movement Behavior Using Web Data for Context Enrichment. *In*: *Pacific Visualization Symposium (PacificVis), 2014 IEEE*. IEEE, 193–200.

Kurashima, T., Iwata, T., Irie, G., and Fujimura, K., 2013. Travel route recommendation using geotagged photos. *Knowledge and Information Systems*, 37 (1), 37–60.

Lee, R. and Sumiya, K., 2010. Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. *In*: *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*. ACM, 1–10.

Li, M., Sagl, G., Mburu, L., and Fan, H., 2016. A contextualized and personalized model to predict user interest using location-based social networks. *Computers, Environment and Urban Systems*, 58, 97–106.

Li, M., Sun, Y., and Fan, H., 2015. Contextualized Relevance Evaluation of Geographic Information for Mobile Users in Location-Based Social Networks. *ISPRS International Journal of Geo-Information*, 4 (2), 799–814.

Liu, X., Liu, Y., Aberer, K., and Miao, C., 2013. Personalized point-of-interest recommendation by mining users' preference transition. *In*: *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. San Francisco, CA, USA, 733–738.

Liu, X., Troncy, R., and Huet, B., 2011. Using social media to identify events. *In*: *Proceedings of the 3rd ACM SIGMM international workshop on Social media*. 3–8.

Liu, Y., Sui, Z., Kang, C., and Gao, Y., 2014. Uncovering Patterns of Inter-Urban Trip and Spatial Interaction from Social Media Check-In Data. *PLoS ONE*, 9 (1), e86026.

McKenzie, G., Adams, B., and Janowicz, K., 2013. A Thematic Approach to User Similarity Built on Geosocial Check-ins. *In*: D. Vandenbroucke, B. Bucher, and J. Crompvoets, eds. *Geographic Information Science at the Heart of Europe*. Cham: Springer International Publishing, 39–53.

McKenzie, G., Janowicz, K., Gao, S., and Gong, L., 2015. How where is when? On the regional variability and resolution of geosocial temporal signatures for points of interest. *Computers, Environment and Urban Systems*, 54, 336–346.

Miller, G.A., 1955. Note on the bias of information estimates. *In*: *Information Theory in Psychology: Problems and Methods*. Free Press, 95–100.

Noulas, A. and Scellato, S., 2012. Mining user mobility features for next place prediction in location-

based services. *In*: *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. Brussels, Belgium, 1038–1043.

Noulas, A., Scellato, S., Lathia, N., and Mascolo, C., 2012. A Random Walk around the City: New Venue Recommendation in Location-Based Social Networks. *In*: *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. Ieee, 144–153.

Parent, C., Spaccapietra, S., Renso, C., Andrienko, G., Andrienko, N., Bogorny, V., Damiani, M.L., Gkoulalas-Divanis, A., Macedo, J., Pelekis, N., Theodoridis, Y., and Yan, Z., 2013. Semantic trajectories modeling and analysis. *ACM Computing Surveys*, 45 (4), 42:1–42:32.

Pham, M.C., Cao, Y., Klamma, R., and Jarke, M., 2011. A Clustering Approach for Collaborative Filtering Recommendation Using Social Network Analysis. *Journal of Universal Computer Science*, 17 (4), 583–604.

Preoţiuc-Pietro, D. and Cohn, T., 2013. Mining user behaviours: A study of check-in patterns in Location Based Social Netoworks. *In*: *Proceedings of the 5th Annual ACM Web Science Conference on - WebSci '13*. New York, New York, USA: ACM Press, 306–315.

Quercia, D. and Lathia, N., 2010. Recommending social events from mobile phone location data. *In*: *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. Sydney, Australia, 971–976.

Rattenbury, T., Good, N., and Naaman, M., 2007. Towards automatic extraction of event and place semantics from flickr tags. *In*: *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 103–110.

Ruths, D. and Pfeffer, J., 2014. Social media for large studies of behavior. *Science*, 346 (6213), 1063–1064.

Salah, A.A., Gevers, T., Sebe, N., and Vinciarelli, A., eds., 2010. *Human Behavior Understanding*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Sengstock, C., Gertz, M., Flatow, F., and Abdelhaq, H., 2013. A probablistic model for spatio-temporal signal extraction from social media. *In*: *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL'13*. New York, New York, USA: ACM Press, 264–273.

Shannon, C.E., 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27 (3), 379–423.

Sheth, A., 2009. Citizen Sensing, Social Signals, and Enriching Human Experience. *IEEE Internet Computing*, 13 (4), 87–92.

Song, C., Qu, Z., Blumm, N., and Barabasi, A.-L., 2010. Limits of Predictability in Human Mobility. *Science*, 327 (5968), 1018–1021.

Steiger, E., Westerholt, R., Resch, B., and Zipf, A., 2015. Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. *Computers, Environment and Urban Systems*, 54, 255–265.

Sun, Y., Fan, H., Li, M., and Zipf, A., 2016. Identifying the city center using human travel flows generated from location-based social networking data. *Environment and Planning B: Planning and Design*, 43 (3), 480–498.

Wilson, E., 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22 (158), 209–212.

Woerndl, W., Brocco, M., and Eigner, R., 2009. Context-Aware Recommender Systems in Mobile Scenarios. *International Journal of Information Technology and Web Engineering*, 4 (1), 67–85.

Ye, M., Yin, P., Lee, W.-C., and Lee, D.-L., 2011. Exploiting geographical influence for collaborative point-of-interest recommendation. *In*: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11*. New York, New York, USA: ACM Press, 325.