Dissecting Foursquare Venue Popularity via Random Region Sampling*

Yanhua Li†, Moritz Steiner§, Limin Wang§, Zhi-Li Zhang† and Jie Bao†
†Dept. of Computer Science & Engineering, Univ. of Minnesota, Twin Cities, Minneapolis, MN, US
§Bell Labs, Murray Hill, NJ, US
{yanhua,zhzhang,baojie}@cs.umn.edu,{moritz,liminwang}@bell-labs.com

ABSTRACT

Location based social networks (LBSNs) are becoming increasingly popular with the fast deployment of broadband mobile networks and the growing prevalence of versatile mobile devices. This success has attracted great interest in studying and measuring the characteristics of LBSNs. However, it is often prohibitive, and sometimes impossible, to obtain a detailed and complete snapshot of a LBSN due to its usually massive scale and the lack of proper tools. In this work, we focus on sampling and estimating restricted geographic regions in LBSNs, such as cities or states, in Foursquare. By utilizing the geographic search APIs provided by Foursquare, we propose a random region sampling algorithm that allows us to draw representative samples of venues (i.e., places), and design unbiased estimators of regional characteristics of venues. Moreover, using a unique dataset with 2.4 million venues, that we collected from Foursquare, we further explore the factors affecting the venue popularity, and present our preliminary findings, with applications in venue recommendation and advertising in LBSNs.

Categories and Subject Descriptors

H.3.5 [**Information Systems Applications**]: Information storage and retrieval—*Online Information Services*; H.2.8 [**Information System**]: Database management—*Database Applications*

Keywords

Location based social networks, sampling, Foursquare

1. INTRODUCTION

The fast development of broadband mobile networks and the increasing prevalence of versatile mobile devices, e.g., smart phones and tablets, help to boost the popularity of location based services. For example, Foursquare [1], one of the most popular location based social networks (LBSNs), had more than 20 million registered users with 2 billion check-ins by April 2012 [3].

This success of LBSNs has generated great interest in studying and measuring their characteristics [6, 4]. By collecting and investigating large scale datasets, these studies provide useful insights into the understanding of different aspects of LBSNs, such as popular route discovery and user mobility prediction. All these efforts rely on mining and understanding massive location-based social network data, and the representativeness (or biasedness) of the dataset may significantly impact the results. However, exhaustive search is in general costly, thus can only be applied to relatively small regions. On the other hand, sampling [5, 7] is a more efficient and practical alternative to timely learn and estimate the statistics of LBSNs, i.e., the total number of venues, check-in distributions, etc. No work so far has been done to systematically study how to efficiently sample a representative venue set from a restricted region, e.g., New York City, in LBSNs.

Contributions. We make the first attempt to study sampling and estimation in LBSNs, and aim to obtain a representative sample set of venues (i.e., places) in a restricted region from a LBSN. Taking Foursquare as an example, we develop a random region sampling algorithm and unbiased estimators for estimating various venue statistics. By analyzing the venue set collected, we explore and understand the key factors affecting venue popularity, with applications in venue recommendation and advertising in LBSNs.

2. BASICS OF FOURSQUARE

Foursquare [1] is a location-based social networking website created in 2009, and has become one of the most popular LBSNs. A Foursquare venue is a physical location. It can be a place of business office or private residence where Foursquare users can check in, e.g., a restaurant, train station or movie theater. Foursquare users can create venues, by providing a few venue attributes, such as the venue's location, name, address, and category, zip code, cross street, phone number, etc. Foursquare allows registered users to explicitly post their presence at a venue, and leave tips to venues for other users to read. These tips serve as suggestions for great things to do, see or eat at the location.

Foursquare venue search API. Given a bounding box of a geographic region, specified by the south-west and north-east locations, the *Foursquare venue search API* [2] returns a list of venues in that region, which gives access to venues and rich information about them, such as venue profile information as well as venue statistics, e.g., the number of check-ins the venue had. However, the API suffers from its space constraint, i.e., up to $\bar{b} = 50$ venues can be returned by a query, and its time constraint, i.e., up to 500 queries can be performed per hour per authenticated account, which significantly limit the speed of venue collection.

Exhaustive venue search. The space limitation of the Foursquare venue search API implies that if an API query gets fewer than 50 venues, the region has been exhaustively explored, providing us a good criterion to perform an exhaustive search for any geographic region. We can divide any objective geographic region into small bounding boxes, so queries performed on each box return fewer than 50 venues.

*This study was done partly while Yanhua Li was a summer intern at Bell Labs, Alcatel-Lucent. It was supported in part by the US NSF grant 0831734, CNS-1017647, the DTRA grant HDTRA1-09-1-0050, and a DoD ARO MURI Award W911NF-12-1-0385.

CoNEXT Student'12, December 10, 2012, Nice, France.

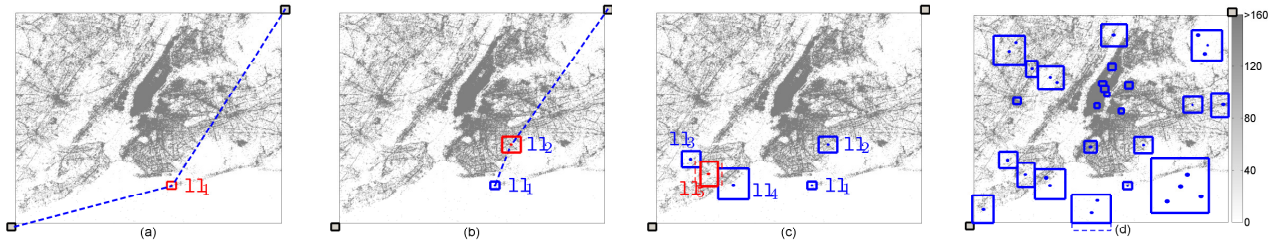

Figure 1: Illustration of the RRS algorithm. The background color represents venue densities as shown on the colorbar. (a) Two starting points $\texttt{ll}_{\texttt{sw}}$ and $\texttt{ll}_{\texttt{ne}}$ are assigned with an initial venue density d_0. Hence, the predicted density for $\texttt{ll}_1$ is $d'_1 = d_0$, and the side length s_1 satisfies $s_1^2 = \bar{b}/d'_1$. By exhaustively searching this area, we learn its true venue density d_1. (b) $\texttt{ll}_2$ is closer to $\texttt{ll}_1$ and $\texttt{ll}_{ne}$. Its density is predicted as the weighted average of d_1 and d_0, with weights as the inverse of its distances to $\texttt{ll}_1$ and $\texttt{ll}_{ne}$. (c) $\texttt{ll}_5$ is chosen between two nearby sampled boxes, and its box is cut to be non-overlapping with them. (d) Over steps, a sequence of sample boxes are formed with more and more accurate densities predicted. Locations falling into already sampled boxes trigger the corresponding boxes being re-sampled.

However, it is very costly to perform such an exhaustive search on a large and dense geographic region, for example, New York City, where the dimension of small searching boxes needs to be within a few meters to meet the criterion. Hence, this method is not suitable for timely learning and estimating the statistics of LBSNs. This motivates us to design efficient sampling methods to collect a representative venue set from the objective region and design unbiased estimators for estimating the statistics such as the total number of venues and venue label distributions.

3. RANDOM REGION SAMPLING

The basic idea behind the random region sampling (RRS) is that given a sampling budget B, i.e., the number of queries allowed, at every step, a location, $\texttt{ll}$, is chosen uniformly at random from the objective region G, and the size of the sample box around $\texttt{ll}$ is determined by two criteria as follows.

• **Box size selection using venue density prediction:** The venue density d' around $\texttt{ll}$ is predicted as the weighted average of the venue densities of its closest sampled boxes. The side length s of a new box centered at $\texttt{ll}$ is computed as $\sqrt{\bar{b}/d'}$, to keep the expected number of venues in this new box close to the API return limit $\bar{b}$.

• **Non-overlapping boxes:** Check whether the new box obtained above collides with any already sampled box, and cut the new box if necessary to keep it containing $\texttt{ll}$ and non-overlapping with those sampled boxes.

Using the above two criteria, a non-overlapping new box $G_{\texttt{ll}}$ is determined based on the best knowledge to have an expected total number of venues close to the API return limit. Then, an exhaustive search is performed on this box, which cumulates one true venue density in G, thus improves the following venue density predictions. The area it covers is considered as a sampled box. Later if a random location is chosen that falls into this box, it is considered that this box is sampled again, and no actual API queries are needed. Note that keeping the sampled boxes to be non-overlapping ensures each box having an invariant probability to be sampled again once the box has been established in one sampling process. This probability is proportional to the size of its area, and it is an important quantity in designing unbiased estimators for the objective region. Until running out of the API budget B, m samples, $X_1, \cdots, X_m$ are drawn from n non-overlapped boxes $G_1, \cdots, G_n$, with $m \geq n$. If the budget runs out while exhaustively searching the last box, that box will be ignored. Fig 1 takes New York City as an example to illustrate how RRS works.

Estimators. Theorem 1 below presents how to use the venues sampled with RRS to estimate the total number of venues N of objective region G, by the estimator $\hat{N}$, where we omit the proof and the evaluation results due to the space limit.

THEOREM 1. *Using RRS with budget B, we obtain m sampled boxes $X_1, \cdots, X_m$. Let $a(X_t)$ and $f(X_t)$ be the size and number of venues of X_t, $1 \leq t \leq m$. Then, $\hat{N}$ in eq.(1) is an (asymptotically) unbiased estimator of N.*

$$\hat{N} = \frac{a(G)}{m} \sum_{t=1}^{m} \frac{f(X_t)}{a(X_t)}. \quad (1)$$

4. EXPLORING VENUE POPULARITY

Now, we investigate how various factors affect the venue popularity in Foursquare, by analyzing a unique dataset with 2,398,931 venues collected from 14 regions during 05/01/12–06/30/12, including New York City, Paris, Seoul, covering a wide range of geographic areas. Our main results are summarized as follows:

• **Venue profile.** Venues with more complete profile information are more likely to be popular, and the two most influential individual attributes are "contact" and "cross street".

• **Venue category.** By performing comprehensive categorical analysis, we observe that venues in the Food category attract the most (43%) public comments (tips) by users, and the Travel & Transport category is the most popular category with the highest per venue check-ins. The residence, office, and school have the highest *user stickiness*, i.e., the average number of repeated visits of users to each venue.

• **Venue age.** The most popular venues were usually created at the early phase of Foursquare.

5. CONCLUSION

Taking Foursquare as an example, we study how to sample and estimate characteristics of location based social networks. Moreover, the findings of our venue popularity analysis may help advertisers to select promising candidate venues for effective advertisement placement, and venue owners to improve venues' attraction to customers.

6. REFERENCES

[1] Foursquare. *https://foursquare.com/*.

[2] Foursquare API. *https://developer.foursquare.com/*.

[3] Foursquare on wiki. *http://en.wikipedia.org/wiki/Foursquare*.

[4] Jie Bao, Yu Zheng, and Mohamed F. Mokbel. Location-based and preference-aware recommendation using sparse geo-social networking data. In *ACM GIS*, 2012.

[5] Abedelaziz Mohaisen, Pengkui Luo, Yanhua Li, Yongdae Kim, and Zhi-Li Zhang. Measuring bias in the mixing time of social graphs due to graph sampling. In *Milcom*, 2012.

[6] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. An empirical study of geographic user activity patterns in foursquare. *ICWSM'11*.

[7] Jia Zhou, Yanhua Li, Vijay Kumar Adhikari, and Zhi-Li Zhang. Counting youtube videos via random prefix sampling. In *IMC*, 2011.