An Empirical Study of Geographic User Activity Patterns in Foursquare

Anastasios Noulas
Computer Laboratory
University of Cambridge
anastasios.noulas@cl.cam.ac.uk

Salvatore Scellato
Computer Laboratory
University of Cambridge
salvatore.scellato@cl.cam.ac.uk

Cecilia Mascolo
Computer Laboratory
University of Cambridge
cecilia.mascolo@cl.cam.ac.uk

Massimiliano Pontil
Computer Science Department
University College London
m.pontil@cs.ucl.ac.uk

Abstract

We present a large-scale study of user behavior in Foursquare, conducted on a dataset of about 700 thousand users that spans a period of more than 100 days. We analyze user checkin dynamics, demonstrating how it reveals meaningful spatio-temporal patterns and offers the opportunity to study both user mobility and urban spaces. Our aim is to inform on how scientific researchers could utilise data generated in Location-based Social Networks to attain a deeper understanding of human mobility and how developers may take advantage of such systems to enhance applications such as recommender systems.

Introduction

During the past decade the widespread use of mobile phones has offered the opportunity to gain insights on human mobility at unprecedented temporal and user participation scales, using Bluetooth or cellular data (Eagle and Pentland 2006; Gonzalez, Hidalgo, and Barabasi 2008). The rise of online social networks and services has provided another useful source where location data and human activity or relationships are being described.

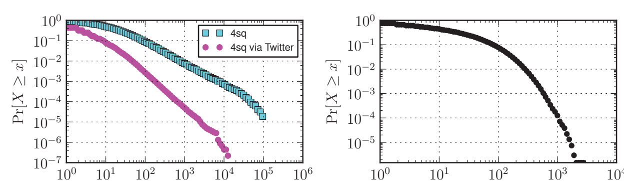
The latter has so far been exploited to both address classical problems arising in social networks and propose directions for new applications. In particular, inferring friendships in a social network through the exploitation of location information such as user geographic co-occurrences has been suggested (Crandall et al. 2010; Eagle, Pentland, and Lazer 2009). The authors in (Cranshaw et al. 2010) improve link prediction based on co-occurrences by taking into account contextual information such as the entropy of a location, a metric that accounts for the diversity of unique visitors to an area. The complementary task, i.e. predicting a person's location based on where her friends are has also been tackled (Backstrom, Sun, and Marlow 2010).

In the meantime, Online Location-based Social Networks (LBSN), such as Brightkite, Foursquare and Gowalla, are services that are built upon the notion of bringing together the places we visit with the friends we connect to, and due to their growing popularity they present a promising source of human activity data.

In this work we present the first large-scale study of user behavior on the most popular LBSN, Foursquare. We have collected approximately 12,000,000 user *checkins* over a period of 111 days, describing the mobility patterns of more than 679,000 users across about 3 million geo-tagged and categorized venues. Our dataset corresponds to a large sample of user activity on Foursquare with fine-grained temporal and spatial information. We present an *analysis of the geo-temporal dynamics of collective user activity* on Foursquare and show how checkins provide a means to uncover human daily and weekly patterns, urban neighborhood properties and recurrent transitions between different activities.

Our results provide strong indications that LBSNs present exciting and promising research opportunities. The application potential of this analysis is remarkable and ranges from more precise location/activity recommender systems and trip advisors to more general fields including urban planning and social sciences.

Foursquare Dataset



(a) Place Checkins CCDF (b) User Checkins CCDF

Figure 1: Complementary Cumulative Distribution Functions for the number of checkins at places (left) and the number of checkins per user (right).

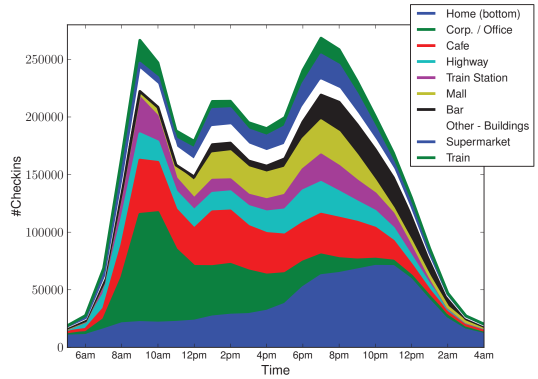
Since Foursquare API provides rate limited authorized access, we have resorted to another channel through which public data is available at large amounts: Twitter messages which contain Foursquare checkins. Through the public stream of Twitter messages, we have recorded approximately 12 million timestamped location checkins, generated by 679 thousand Foursquare users, between May, 27th 2010 and September, 14th 2010. Each collected tweet provides a pointer to the the corresponding venue. Thus, we have requested additional data directly from Foursquare and acquired the following information about a venue: geographic coordinates, category, total number of checkins, unique number of visitors and address. As a result, we have acquired a large set with fine-grained spatial and temporal data. While this methodology allows us to acquire data only about the subset of Foursquare users who have selected to share their checkins publicly via Twitter, our sample represents approximately 20% to 25% of the entire Foursquare user base which amounted in total 3 million users as of September 2010 (Techcrunch 2010).

As an introductory step to our analysis, we present here some comments about how users share information about their locations. The number of checkins is an indicator of popularity for places among users. In Figure 1(a) we report the complementary cumulative distribution function (CCDF) of the number of observed checkins at each location. We present two cases: the number of checkins per location in our dataset, acquired via Twitter over 111 days, and the number of total checkins reported on Foursquare (since its inception) for each venue. While our sample contains only a subset of checkins at each venue, the two distributions exhibit the same trend, with heavy tail and power-law behavior. Only a few places feature a large number of checkins, while a higher number of places have only few checkins. For instance, a central train station will have higher levels of user activity compared to a small park at the outskirts of a city. In Figure 1(b) we show the CCDF of the number of checkins per user. Again, we have a heavy tail in the distribution, denoting how user participation can vary significantly. About 20% of users have just one checkin, with 40% above 10, whereas there is a set of approximately 70,000 active users (around 10%) that has more than 100 checkins. The reasons behind this heterogeneity in user behaviour could be many, ranging from cognitive factors (e.g., if a user forgets checkin at a place), to social ones (e.g., sharing location with friends), as checkins are voluntary and not automatic in Foursquare. Privacy is also a fundamental aspect to consider here.

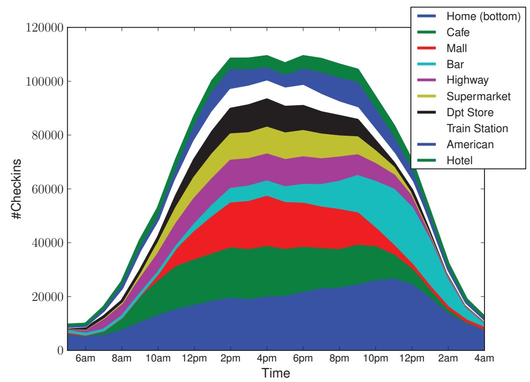
Spatio-Temporal Patterns of User Activity

We now present an analysis of the dynamics of Foursquare user activity. Our findings suggest that activity in Foursquare varies within the course of a day and of a week, with meaningful patterns closely related to human activity from a temporal and spatial point of view. Finally, we describe how checkins constitute spatio-temporal traces that can help to investigate user transitions from one place or activity to the next.

Geo-Temporal Rhythms In Figures 2(a) and 2(b) we plot the number of checkins across the ten most popular categories observed during weekdays and weekends. The two curves present considerably different patterns. At weekdays activity presents three peaks: in the morning when people go to work, at lunchtime and between 6pm and 8pm when they commute, return home or go to malls and bars. On the other hand, during weekends user activity presents a smoother evolution course, reaching a long lasting plateau between 12pm and 10pm. Another difference to note between the two is that category Corporate/Office disappears from the top set of user activities and is substituted by leisure related activities such as *American* (Food) and *Hotel*, while categories such as *Bar* and *Mall* also show increased preference rates among users. In both cases, however, checkins at the *Home* category show a continuous rise throughout the day, with a steeper increase at 6pm during weekdays. While those findings are in line with what may be expected by a human observer, they also demonstrate that these data could be used to record and measure how human communities commit to different tasks over time and benefit research in social sciences or the development of applications for smartphone users.



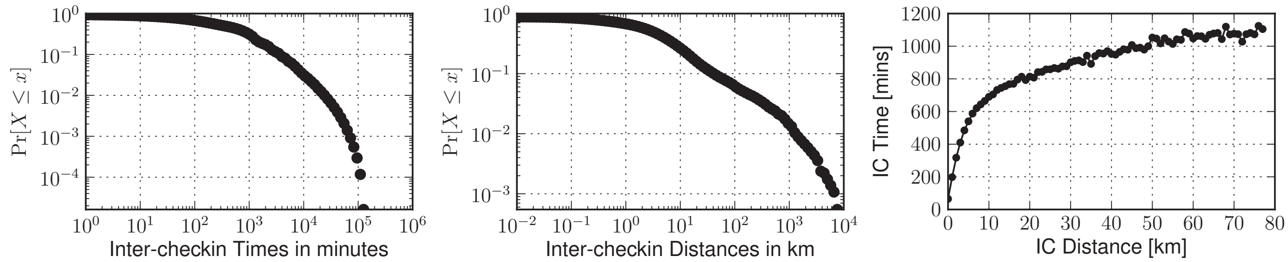
(a) Weekdays



(b) Weekend

Figure 2: Stacked plot of the 10 most popular categories over weekdays and weekends. Popularity decreases bottom-up.

Checkin Dynamics We now investigate how user checkins take place over time and space. In Figure 3(a) we plot the CCDF of *inter-checkin times*, where an inter-checkin time is defined as the temporal interval between two consecutive checkins. More than 10% of checkins occur within 10 minutes, while this fraction rises up to 30% within 100 minutes and a large portion, almost 20%, is longer than 2000 minutes (approximately 33 hours). This represents a major differ-



(a) Inter-Checkin Times CCDF (b) Inter-Checkin Distances CCDF (c) Time Overs Distance

Figure 3: Temporal and spatial intervals of consecutive user checkins (a),(b) and median inter-checkin time over distance (c).

ence with previous studies (Zheng et al. 2008), where user location was periodically sampled through GPS. As a consequence, the frequency at which users report their locations in the system should be taken into account in every specific usage scenario, either for scientific research or for application design. For example, two temporally close checkins of the same user could signal an important correlation between two locations, but as this temporal distance increases we could express higher confidence that the two checkins are not strictly consequential.

Similarly, in Figure 3(b) we plot the CCDF of *inter-checkin distances*. It exhibits a comparable power law behaviour: 20% of checkins occur within a distance of 1 km, a significant proportion of 60% between 1 and 10 km, 20% of checkins take place at distances over 10 km and a small portion of these, around 5%, extend to distances beyond 100 km. However, it is likely that longer inter-checking distances are correlated with longer inter-checkin times, since physical motion between distant locations inevitably requires longer periods of time.

In fact, we demonstrate how inter-checkin times and inter-checkin distances are correlated in Figure 3(c), where we plot the median inter-checkin time as a function of the inter-checkin distance. For each pair of subsequent checkins we have rounded the distance between the locations involved with precision of plus/minus 500 metres and then measured the median value of the distribution of checkins that took place across a given distance. Inter-checkin times observed within 500 metres have a median value of approximately one hour. The trend of observing a larger temporal interval as users cover larger distances increases fast for distances within a range of 6 kilometres, whereas above this threshold the rate of increase slows significantly and stabilizes at relatively large distances of 60 to 70 kilometres.

Activity Transitions An important aspect to consider with respect to human behaviour, is how different activities succeed each other. The question we ask is: *If person X is engaged in activity Y (i.e., visiting a place belonging to category Y), which activity will follow next?* In particular, we calculate the transition probability $P_t(i,j)$ from category i to category j as:

$$P_t(i,j) = \frac{c_{ij}}{\sum_{k \in C} c_{ik}} \quad (1)$$

Category	Category	P_t^{10}	P_t^*
Train	Train Station	0.48	0.30
Terminal	Airport	0.46	0.17
Gate	Airport	0.45	0.22
Moroccan	Theme Park	0.39	0.06
Train Station	Train	0.38	0.22
Rental Car	Airport	0.36	0.18
Plane / In-flight	Airport	0.33	0.19
Tram	Airport	0.33	0.19
Cineplex	Mall	0.30	0.08
Plane	Airport	0.28	0.15
Bridge	Highway / Traffic	0.28	0.10
Lab	University	0.26	0.09
Surf Spot	Beach	0.25	0.06
Trade/Tech School	Other - Buildings	0.25	0.07
Emergency Room	Hospital	0.25	0.08
Hotel Bar	Hotel	0.24	0.07
Engineering	University	0.24	0.07
Movie Theater	Mall	0.24	0.06
Other - Travel	Highway / Traffic	0.23	0.11
Taxi	Highway / Traffic	0.23	0.09

Table 1: Top-20 Activity Transition Probabilities for P_t^{10}.

where C denotes the set of all categories and c_{ij} the number of checkin transitions from a place in category i to one in category j. We let P_t^{10} be the transition probability between consecutive categories conditioned on the transition time being less than 10 minutes. We impose this threshold in order to apply a *mobility* bias to the transition data. In Table 1 the top twenty transition probability values between categories are reported. As a comparison we list the values of P_t^*, which has no temporal threshold imposed. These probabilities could be interpreted as weights of transitions on an *activity* graph G, where nodes are the 288 place categories and edges exist if transitions have occurred from one category to another. Comparing P_t^{10} and P_t^*, we notice that probability P_t^* is much lower than P_t^{10}: when calculating transitions without considering the time between subsequent checkins then the *activity* transition graph becomes *noisy*, as users checkin voluntarily and certain transitions may not be reported. The intriguing aspect of the probabilities calculated in Table 1 is that the LBSN user generated data can provide valuable insights on how activities of mobile users success each other. Moreover, those activity networks could

Category	Category	Dist(km)	checkins	Category	Category	Dist(km)	checkins
Train Station	Train Station	3.5	231	Airport	Gate	543.7	147
Train Station	Technology	0.16	187	Airport	Airport	1177	107
Light Rail	Mall	0.02	122	Airport	Airport	1092	71
Train Station	Monument / Landmark	2.0	101	Monument / Landmark	Train Station	267	65
Train Station	Convention Center	0.57	97	Parks & Outdoor	Theme Park	0.37	62
Mall	Movie Theater	0.01	73	Gym	Hospital	2.18	47
Monument / Landmark	Park	4.34	67	Coworking Space	Coffee Shop	0.04	47
Mall	Apparel	0.50	64	Airport	Gate	3973.3	45
Corporate / Office	Corporate / Office	0.60	61	Train Station	Bus Station	329	45
Playground	Karaoke	0.50	59	Gym	Corporate / Office	2.09	44

Table 2: Top Place Transitions for inter-checkin times of 0-10 minute (left) and 100-500 (right) minute intervals.

be calculated for different temporal periods (i.e. morning, night etc.) or geographic regions (Europe, US, Asia etc.). While mobile applications that target content to mobile users could be profoundly benefited from such analysis, we now demonstrate how the possibility for location-based marketing or other strategies also rises.

Place Transitions In Table 2 we present the top ten place transitions in terms of checkin frequency for inter-checkin times that fall in the 0 to 10 minute (left) and 100 to 500 minute intervals (right). To facilitate interpretation we show the categories of those places, whereas the distance and the number of checkins between the two locations are also reported. We can see that that transitions occurring at smaller intervals concern places that are geographically proximate. Such cases could include travels between train stations, leaving the mall to watch a movie or the office to attend a meeting nearby. Switching our view to the 100 to 500 minute interval, the analysis reveals a new class of place/activity transitions. While people fly from one city to another and checkin at the corresponding airports, they unfold spatio-temporal connections of larger scales. The most frequent location transition in this case is between San Francisco International Airport and Los Angeles International Airport (LAX). The eighth transition with a distance of 3973.3 km is that between LAX and JFK Airport in New York. Another observation is the fifth case that takes place within Disneyland at Anaheim, California. Despite occurring within a small spatial interval, this may indicate an existence of a location where people stay for a long duration (for instance a restaurant) and then move to a different place nearby. The analysis of the place and activity networks in Foursquare opens a new direction for research, since for the first time global and regional transport networks can be uncovered, while a deeper understanding of collective human activity could also be acquired.

Discussion and Future Work

LBSN data offers unprecedented opportunities. Never before have such large communities of individuals been able to exchange data about their location and activities through interfaces that augment the natural environment with new types of data layers. The fact that mobile devices do not act as continuously monitoring sensors, but instead represent opportunistic gateways that users can exploit at will, poses challenges at the analysis and interpretation of the data. As it has been shown, imposing spatio-temporal thresholds can unfold different classes of place transitions or help in the identification of sequential activity transitions.

Moreover, the analysis of the geo-temporal rhythms of user checkins informs us on the general consensus of user activity at a given time and place. A recommendation or advertising application could take this into account in order to tune its content. The activity and place transitions previously analyzed could also contribute to this direction.

In terms of future work we intend to study the development of prediction frameworks on user activity and mobility. Moreover, a problem not addressed in this paper, is how users in LBSNs generate content, such as comments, tips and tags. Topic modelling techniques could be used in order to extract information through textual data overlaid on the geo-social plane.

References

Backstrom, L.; Sun, E.; and Marlow, C. 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of WWW'10*.

Crandall, D. J.; Backstrom, L.; Cosley, D.; Suri, S.; Huttenlocher, D.; and Kleinberg, J. 2010. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences* 107(52):22436–22441.

Cranshaw, J.; Toch, E.; Hong, J.; Kittur, A.; and Sadeh, N. 2010. Bridging the gap between physical location and online social networks. In *Proceedings of Ubicomp '10*.

Eagle, N., and Pentland, A. 2006. Reality mining: sensing complex social systems. *Personal Ubiquitous Comput.* 10(4):255–268.

Eagle, N.; Pentland, A.; and Lazer, D. 2009. Inferring social network structure using mobile phone data. *Proceedings of the National Academy of Sciences* 106(36):15274–15278.

Gonzalez, M. C.; Hidalgo, C. A.; and Barabasi, A. 2008. Understanding individual human mobility patterns. *Nature* 453(7196):779–782.

Techcrunch. 2010. Foursquare now 3 million strong. http://techcrunch.com/2010/08/29/foursquare-now-3-million-strong/.

Zheng, Y.; Li, Q.; Chen, Y.; Xie, X.; and Ma, W.-Y. 2008. Understanding mobility based on gps data. In *Proceedings of Ubicomp '08*.