Discovering and Predicting User Routines by Differential Analysis of Social Network Traces

Fabio Pianese*, Xueli An*, Fahim Kawsar*, Hiroki Ishizuka†
* Enabling Computing Technologies Domain
Bell Labs, Alcatel-Lucent
Antwerp, Belgium
{fabio.pianese, xueli.an, fahim.kawsar}@alcatel-lucent.com
† University of Tokyo
isi@mcl.iis.u-tokyo.ac.jp

***Abstract*—The study of human activity patterns traditionally relies on the continuous tracking of user location. We approach the problem of activity pattern discovery from a new perspective which is rapidly gaining attention. Instead of actively sampling increasing volumes of sensor data, we explore the participatory sensing potential of multiple mobile social networks, on which users often disclose information about their location and the venues they visit. In this paper, we present automated techniques for filtering, aggregating, and processing combined social networking traces with the goal of extracting descriptions of regularly-occurring user activities, which we refer to as "user routines". We report our findings based on two localized data sets about a single pool of users: the former contains public geo-tagged Twitter messages, the latter Foursquare check-ins that provide us with meaningful venue information about the locations we observe. We analyze and combine the two datasets to highlight their properties and show how the emergent features can enhance our understanding of users' daily schedule. Finally, we evaluate and discuss the potential of routine descriptions for predicting future user activity and location.**

I. Introduction

The commercial availability of personal mobile devices capable of location sensing has stimulated research on techniques to analyze, explain, and predict geographical mobility and other aspects of user behavior. Current smartphones models are all capable of continuously running data collection software to aggregate a number of sensor readings. These devices can thus provide a faithful record of a user's mobility. By collecting GPS coordinates, SSIDs of nearby WiFi access points and their signal strength, GSM/UMTS cell ID strings, Bluetooth proximity information, accelerometer readings, etc. it is possible to piece together an increasingly detailed picture of user habits, spatial trajectories, and mutual interactions.

A number of studies based on the continuous collection of location data attempt to infer user activity from a combination of temporal aspects (such as time of day or duration information about a user's presence in a certain location), frequency and distribution of visits over time (such as whether they happen on weekdays or weekends), and approximate matching of location data with external sources of mapping information [1]–[8]. Despite the good spatial accuracy that can be achieved with location sensing techniques, the major issue with their interpretation remains the *lack of contextual and semantic information* to help reconstruct the purpose and significance of a user being in a given location at a given time. User intervention is often required to provide the context that cannot be inferred from the location traces [9]. Another disadvantage of continuous location tracking is cost: while it is possible to sample sensor data at a high-frequency, the sustained activity of sensing and processing hardware and regular transmission of sensed data may result in both an unacceptable drain on the device's battery life and a waste of network resources.

Social networking mobile applications are increasingly popular among users who voluntarily share many details of their private life: blogging by publishing geo-tagged Twitter[1] messages, advertising their presence in a location with Foursquare[2], looking for nearby friends with Facebook Places[3], etc. These applications leverage the available sensing hardware of the smartphones to provide context-aware services and often include sensor data (mostly in the form of GPS coordinates) alongside user-generated messages. Location data are used by social networks to provide information such as the names of friends in the vicinity, reviews about nearby venues, or announcements about local happenings and public events. Moreover, services such as Foursquare also advertise details about the type and name of the venues a user attends, which is an invaluable piece of information to understand the purpose of a user's presence in a particular location.

In light of the mainstream adoption of context-aware mobile applications, an alternative *participatory approach to user activity sensing* has been gaining traction. Instead of actively generating a continuous stream of sensor readings, it may be more convenient to collect and process the trails of location data that are produced as a side-effect of user activities, in an opportunistic fashion. This emerging approach has a number of benefits and limitations compared to the traditional forms of active tracking, as a consequence of its reliance on some form of user interaction with the monitored infrastructure. On one hand, the data collection is non-intrusive and does not require an always-on software running on the user device,

[1] http://twitter.com/
[2] http://foursquare.com/
[3] http://www.facebook.com/about/location

978-1-4673-5828-6/13/$31.00

which extends battery life, yet it provides meaningful metadata about the venues a user visits; on the other hand, the data points generated could be few and far between, with negative consequences on the accuracy of the tracking. On one hand, users can be informed of the privacy implications of interacting with the monitored infrastructure and can whithold sensitive information by simply avoiding using the system where and when confidentiality is important; on the other hand, important aspects of a user's daily life could remain totally invisible.

In this paper we investigate the following question: *"To what extent can location traces provided by users via normal interactions with social networking applications help us understand and characterize their daily schedule?"* In Section II we present and analyze a data set of localized Twitter messages and Foursquare check-ins, and discuss the limitations of our participatory sensing approach based on observed user behavior. We highlight the remarkable temporal, spatial, and activity type features of the datasets, *introducing the concept of "user routines"* as an approximation of the recurrent patterns in space and time that emerge from the data. Then, in Section III we present a method to extract user routines by constructing event clusters. In Section IV we further refine our understanding of user routines and derive the likely meaning of their features based on Foursquare venue information and other considerations. In Section V we apply a straightforward prediction approach based on the routine information and discuss the results. Finally, Section VI presents the related work and Section VII concludes.

II. Mining Location-Aware Social Networks

Social networks have a great potential as large-scale opportunistic sensing infrastructures. Their users provide data on a *voluntary basis* through *explicit interactions* with social networking software - a method which is non-intrusive and privacy-aware. Privacy concerns are eased by the fact that all data have been explicitly published, leaving users in full control over their public image. Also, collecting location and activity data from social networks does not require privileged access to the user terminal, nor a dedicated and continuous monitoring activity: it is an integral part of the normal interaction between the user, the application, and the network without additional overhead. Furthermore, a social network can facilitate the matching between user-generated data and high-quality semantic information, often revealing the meaning and purpose of the presence of a user in a given location. It is an interesting question whether the normal interaction with social networks can become a suitable source of insights on user behavior. In this section we explore the trade-offs between accuracy, which is challenged because of the nature of user interactions with social applications, and expressivity, which is enhanced by the features of the social platforms and the modes of interaction users have with them.

A. Collecting Traces: Methodology and Characteristics

We constructed our dataset by mining public user traces generated by the Foursquare social network using a side-channel approach to data collection. As a large number of Foursquare users also publish their check-in information on the Twitter micro-blogging service via geo-tagged *tweet* messages, we crawl Twitter and collect the message history for users who routinely publish their Foursquare data. The Foursquare check-in tweets we collect contain a human-readable message, the GPS co-ordinates of the current user location, and a URL pointing to the relevant Foursquare information. We then extract Foursquare metadata that identifies the location of the check-in by accessing the URL, and finally retrieve the information about the venue (including name and category of the venue) from its Foursquare web page.

B. Foursquare Trace Extraction and Pre-processing

For this study, we monitored the geo-tagged tweets broadcast by 14,587 users in the Tokyo Metropolitan area, defined as a circle with a 30 Km radius from the center of Tokyo, over eleven months from end July 2010 to the beginning of July 2011. During this period, we collected a total of 179,372 geo-tagged tweets from the ***observed users*** inside our region of interest. Out of these tweets, a large amount (about 50%) had embedded Foursquare check-in information. At preliminary inspection, we observed a strong variability in user check-in behavior: few users were very active, with occasional large bursts of messages on a same day, while most users had a negligible overall activity.

We then singled out the users who published Twitter messages carrying a Foursquare check-in with *an average of at least one message per week during the entire observation period*. This corresponds in our case to an overall threshold of 48 messages, which we consider a minimum condition for users to be regularly accessing the Foursquare application. We deliberately avoid introducing further limiting criteria, such as conditions on the distribution over time of user-generated events, in order to lend further credibility to our analysis. Users who exceed the minimal level of activity, whether in a steady pattern or in a few isolated bursts, are thus both included in the resulting dataset. After this filtering process, we obtain a set of 825 ***active users*** (5.66% of the initial amount of users). Active users generated a total of 157,806 geo-tagged tweets (87.98% of the observed total) out of which 79,341 turn out to contain Foursquare check-ins. Overall, we noticed that on most days, individual users either do not generate on average any Foursquare events (82.5% of the days) or just produce a singleton (12.0% of the days). Only in the remaining 5.5% of the days users generate two or more check-ins.

After making sure that the coordinates from the GPS data in the geo-tagged tweets agree with the corresponding Foursquare location coordinates, we generate two separate traces: one with the Foursquare check-in data (4SQ), one with the remaining geo-tagged tweets (GTW). We further partition each of our complete (ALL) datasets into a weekday (WD) and a weekend and holiday (WE) trace. At the end of these pre-processing steps, our 4SQ-WD dataset includes 55,473 check-ins, while the 4SQ-WE contains 23,958. Our GTW-WD dataset contains 110,483 geo-tagged tweets, while the GTW-WE contains

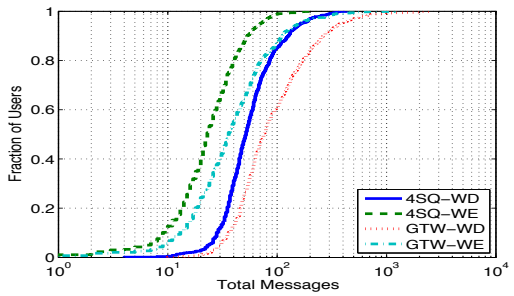

Figure 1. Cumulative Distribution of Total Observed Events per User

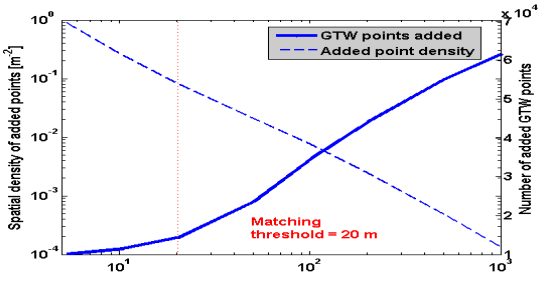

Figure 2. Number of GTW events associated to venues in the 4SQ trace for a range of matching distance threshold values shown on the horizontal axis

46,824 events. In Figure 1 we plot the empirical cumulative distribution of the users by total number of messages generated during the entire observed period.

We observe that the average daily generation rate of Foursquare check-ins in WE traces is only slightly higher compared to WD traces. While the average number of weekly check-ins for 4SQ users is about 1.4 (WD) and 0.6 (WE), the medians of the same datasets are about 0.9 and 0.4, respectively. This further confirms that even among the active users the event generation behavior is greatly variable, with a few users producing most of the events: the top 10^{th} percentile of users have more than 2.2 and 1.0 average weekly check-ins respectively. As we see in Figure 1, the Foursquare check-in distribution is especially skewed, even in comparison to the corresponding geotagged Twitter event distribution.

C. User Classification: Recurrent Feature Clusters

We now introduce a set of criteria to automatically classify the active users into categories based on the combined presence of recurrent spatial (S), temporal (T), and venue (V) *features* in their event traces. We start by folding the entire observed period into a single-day window that contains all the recorded events with their 24H timestamp, location, and associated venue type information. We then attempt to construct clusters of check-in points independently over the three features: for this task we use the DBSCAN algorithm [10] and empirically define clusters as sets of cardinality equal or greater than $k = 5$ (for WD traces) or $k = 3$ (for WE traces, to account for the lower number of total data points) using three values of a distance (ϵ) parameter:

- For S-type clusters, we use $\epsilon = 20$ meters
- For T-type clusters, we use $\epsilon = 10$ minutes
- For V-type clusters, we use $\epsilon = 0$

In Table I we display the results of the independent clustering operation: the left side of the table presents the breakdown of categories where at least one feature is present; the right side contains the user breakdown by combined features. We can thus notice that clustering by venue is not a discriminating factor for most users (99% in WD traces, 81% in WE traces). Spatial and temporal clusters are more selective: in WD traces, S and T clusters can both be formed for about 65% of users, while in WE traces the rate of success is lower, only 22% and 14% respectively. Interestingly, some users with spatial clusters do not present temporal clusters (about 12% on both WD and WE traces), some users with temporal clusters do not display spatial clusters (14% in WD, 5% in WE), and 19% of users do not have any feature in their WE trace. We attribute the remarkable fraction of users without features in the WE traces both to a smaller overall amount of available data points and to a less structured user behavior during the weekends, when users are likely to be free from the regular structure of work life [11]. Conversely, the STV category *contains users whose trace presents features in all three dimensions*: they amount to 52.5% of the total during workdays and to 9.6% during weekends. In Section V we will focus again on this user category to evaluate our ability to build expectations on future user activity based on emerging feature combinations.

D. Features of the Geo-tagged Twitter Dataset (GTW)

As we did above for our Foursquare dataset, we also determined the properties of the geo-tagged Twitter dataset. In Table II we summarize the temporal and spatial features of GTW traces (which do not contain venue type information). In WD traces we observe spatial and temporal features in about 40% of the users, with both features appearing in 31% of the total and no features in 52%. In WE traces, recurrent features appear to be less common with only 35% of users showing spatial features and 21% temporal features, 17% showing both, and 61% showing none.

E. Augmented Foursquare+Twitter (AFT) Dataset

Finally, we attempt to augment the 4SQ dataset by recognizing events in the GTW dataset that take place in known Foursquare venues. We conjecture that users might often forget to perform check-ins in known venues, a behavior which would affect our ability to reconstruct features from their 4SQ traces. To estimate the extent of this phenomenon, we determine the existence of matches between tagged Foursquare venues and events that appear in the GTW dataset. Our approach is based on spatial inference: we associate tagged check-ins found in our 4SQ dataset with GTW events that were recorded in nearby GPS locations. We use a rather small matching threshold, in order to account (among other factors) for the possible inaccuracies of indoor GPS sensor readings, and conjecture that GTW events happening at such a small

Table I
BREAKDOWN OF ACTIVE USERS IN THE FOURSQUARE CHECK-IN (4SQ) TRACES DISPLAYING FEATURE CLUSTERS
(S = SPATIAL FEATURES, T = TEMPORAL FEATURES, V = VENUE TYPE, * = WILDCARD)

# active users with feature clusters	S*	T*	V*	STV	ST	SV	TV	S	T	V	no clusters
4SQ-WD	539	542	814	433	0	106	108	0	1	167	10
4SQ-WE	183	117	667	79	0	104	38	0	0	446	155
4SQ-ALL	626	664	824	541	0	85	122	0	1	76	0

Table II
BREAKDOWN OF ACTIVE USERS IN THE GEO-TAGGED TWITTER (GTW) TRACES DISPLAYING FEATURE CLUSTERS
(S = SPATIAL FEATURES, T = TEMPORAL FEATURES, * = WILDCARD)

# active users with feature clusters	S*	T*	ST	S	T	no clusters
GTW-WD	291	266	222	69	44	367
GTW-WE	218	129	105	113	24	385
GTW-ALL	319	302	252	67	50	360

distance were generated during a visit of the known Foursquare location. Once a match between a GTW point and an existing 4SQ check-in is found, the GTW event is tagged with the same venue type information as the matching check-in and added into an *Augmented Foursquare + Twitter (AFT)* dataset, together with all the Foursquare check-ins contained in the 4SQ dataset.

Figure 2 represents the amount of points that our approach could match as a function of the threshold distance value, and shows at the same time the variation in the spatial density of the added points. Using a sensibly small 20 meter threshold we manage to label with inferred semantic information about 14,000 additional data points from the GTW traces of 463 users. The matched events amount in size to more than 20% of the original 4SQ dataset. The individual user with the largest amount of matches gained 880 new events. However, the median amount of matched events per user was 6, and only 170 users received more than 10 events each. We observe that the addition of the matched GTW data leads to the emergence of previously hidden spatial features only in about 5% (WD) and 3% (WE) of the 4SQ user population, with only a slight reduction in the number of the WE users without detected features (-2% of the total users). We observe that trace augmentation by distance-based venue inference is effective at increasing the number of tagged data points, thus considerably improving the sample size. However, we remark that this technique has a minor role in discovering previously undetected features in our data, and conclude that the occasional skipping of venue check-ins by regular Foursquare users does not adversely impact our ability to detect features in their traces.

III. ON USER ROUTINES

Data obtained from participatory sensing sources can be problematic to work with. For instance, the Foursquare trace we presented in the previous Section is remarkable for the scarcity of data points and for their irregular distribution over time and across users. The dearth of subsequent check-ins during a same day rules out the use of probabilistic models to capture the transitions between subsequent locations, an approach to mobility detection which has been attempted with some success on datasets of much larger entity without constraints of geographic scope [11]. Furthermore, we lack explicit information on the dwelling time of a user in each location, a fundamental aspect of time series analysis [8]. These limitations encourage us to adopt a different approach that focuses on recurrent patterns of events with common characteristics, rather than attempting to reconstruct a likely sequence of daily transitions between events. The dataset features we used in our initial analysis isolate groups of user-generated events that have common aspects and that repeat in multiple occasions. However, features must be present over multiple days in order to be qualified as important in the daily schedule of a user.

We now introduce **user routines** as a model for regular user activities. A user routine can be defined as the *repeated occurrence of similar features that happen at the same approximate time of day on a number of days*. Ideally, user routines capture purposeful repeated visits to important locations, lifestyle habits, and regular behaviors that occur often enough to represent a relevant part of a user's daily life. Collected information about events over time also allows us to estimate users' typical dwelling time in a given situation. In this section, we present an automated method of routine discovery based on combined clustering steps and empirically evaluate its sensitivity to parametric choices using our datasets.

A. Multi-Feature Clusters for Routine Detection

To characterize salient aspects of user routines, we will consider the distribution over time of user reported events in conjunction with some of the event properties. Intuitively, a visit to a bar in the morning (to get breakfast or to buy a snack) is substantially different from a visit to the same bar at night time (to have dinner or to drink with friends); therefore, the event needs to be classified differently in the user's routine, regardless of the fact that both events unfold at the same location. By the same principle, a user who likes to try a different restaurant every night for dinner will perform the same activity at the same time, although the events will be recorded in a multitude of different venues.

1) Temporal Partitions: Individual user activity appears to be unevenly distributed across the day, with characteristic

bursts of events that reflect the user's lifestyle. Based on these features, we consider the temporal dimension as the primary aspect that defines routines. We accordingly introduce a first partitioning of the events on the time domain before further classifying the points based on their features. Temporal partitions are identified using the DBSCAN algorithm [10] with parameters $k = 5$ and $\epsilon = 10$ minutes; the centroids of each cluster are calculated by averaging the time coordinates of the events contained; the boundaries of the clusters are subsequently derived by bisecting the interval between each pair of centroids.

2) Spatial-temporal Clusters: For spatial clustering, we utilize again the DBSCAN algorithm over each time interval with parameters k (variable, as we will discuss later in this Section) and $\epsilon = 20$ m. Clusters are further filtered using a *significance criterion* which only validates those clusters that contain points generated over a total timespan of at least k days as relevant to the user's routine. The result is a set of clusters for each time partition that are characterized by the repeated appearance of events in nearby locations over different days.

3) Activity-temporal Clusters: For activity clustering, we consider the occurrence of tagged check-in events over each time partition: clusters are present if at least k events of the same type are available that are generated over a total timespan of at least k days. The result is a set of clusters for each time partition that are characterized by the repeated appearance of venues of a same Foursquare venue type.

B. Sensitivity Analysis and Discussion

Multi-feature cluster generation is sensitive to the choice of parameters provided to the DBSCAN algorithm. While the value of the distance ϵ can be based on empirical considerations such as what constitutes *close* events in the spatial or time domain, the choice of the minimum cluster size k will yield results that depend on less intuitive features of input data that are unique to each user's event trace. To evaluate the consequences of the value of the k parameter on the results, Figure 3 presents a scatterplot of the relation between the number of check-in events and number of spatial-temporal and activity-temporal clusters for 4SQ-WD and 4SQ-WE traces. We expect routines of users with a large amount of check-ins to reflect quite well the actual number of important features of their daily activity, which we also assume to be reasonably small, *e.g.*, not larger than 15 for the most prolific users. Conversely, we expect users with few check-ins to display at least a few clusters. Figure 3 shows that larger values of k enhance the linear dependency between the number of clusters and check-ins, although they reduce the overall observed range of variation. Based on the above considerations, we empirically choose a set of parameters that provide consistent results across most user traces with STV features. In the following, the minimum multi-feature cluster size will be set to $k = 5$ for WD and $k = 3$ for WE traces.

IV. Tagging Activities in User Routines

Previous studies have found that behavior patterns of social network users are heavily dependent on application features and on perceived privacy implications [12]. In this section, we present a method that relies on these differences to better understand routine features detected by multi-feature clustering. We first introduce the concept of Differential Cluster Analysis, a generic technique to combine information from multiple data sources, and then describe heuristics to attribute meaningful labels to clusters in user routines.

Table III
Differential Cluster Analysis: 4SQ and GTW datasets

4SQ + GTW	Week Days (172 users)		Week Ends (47 users)	
Users with:	One	More	One	More
Foursquare Clusters	73	99	34	13
Matching Clusters	20	19	3	3
Invisible Clusters	73	84	32	11
Matching + Invisible	39		6	
GTW events covered	16093		2008	

A. Differential Cluster Analysis

We consider an individual user's spatial-temporal clusters from the 4SQ and GTW datasets and calculate the intersection of the two sets. We assume a match exists across sets if the spatial coordinates of a pair of cluster centroids are close and if their temporal support overlaps by more than half the duration of the shortest cluster of the pair. A spatial tolerance of 50 m is used to account for both lower indoor GPS accuracy and the fact that we are now dealing with clusters, not individual data points. Based on the results of this process, we enumerate three clusters types: Foursquare, Matching, and Invisible. *Foursquare clusters* are found in the 4SQ dataset only, and the events they contain are usually tagged with categories that help us attribute to them an activity meaning. *Matching clusters* are those that exist in both 4SQ and GTW sets; a match across datasets confirms the relevance of a spatial-temporal feature in the daily life of a user and can improve the temporal coverage thanks to the additional data points. The meaning of matching clusters can as well be identified thanks to the tagged events they contain. Finally, *invisible clusters* are those that appear in the GTW dataset but do not match any of the clusters in 4SQ. The presence of invisible clusters provides insights about locations and activities that, although relevant in the user routine, do not show up in a social check-in application trace.

Table III reports the results of the differential clustering technique applied to users of both Foursquare and Geo-tagged Twitter applications who present clusters in both datasets. We can observe that a relatively small number of users (39 WD, 6 WE) have matching clusters between the two datasets: except in the few cases where all GTW clusters match 4SQ clusters (15 users in WD, 4 in WE), users of both applications show two disjoint sets of foursquare and invisible clusters.

Our observations confirm that Foursquare users appear to withhold check-ins when visiting certain locations in which they spend a substantial amount of their time. Recent studies suggest that, among the possible reasons, these venues might be seen as sensitive (home, work) or otherwise not interesting

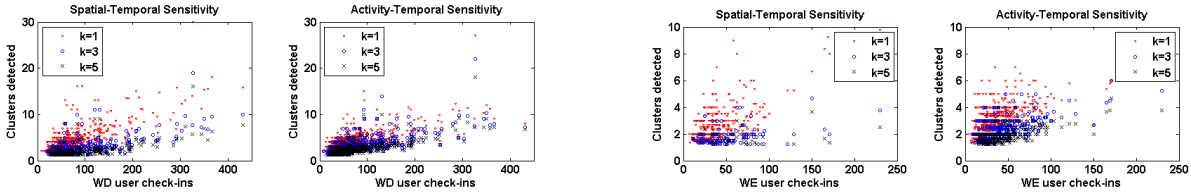

Figure 3. Sensitivity analysis: multi-feature clustering results for $k = 1, 3, 5$ in the 4SQ-WD and 4SQ-WE datasets

to their friends and fellow users [12]. However, it frequently happens that those users disclose their location in messages sent to micro-blogging platforms such as Twitter. Our hypothesis is that users may have a different perception of privacy threats across different social applications or might even be unaware of the inclusion of geo-tagging information in the less social and personal context of a microblogging platform.

Invisible clusters allow us to improve our understanding of user routines. As a qualitative example, in Figure 4 we represent the routine we extracted for a representative user from our dataset. The subject presents several Foursquare clusters, two matching clusters (one in the morning, one in the afternoon), and three invisible clusters. The Foursquare clusters have been identified as Food (breakfast and lunch) and Professional (working at office). Clusters labeled 1 to 4 in the picture are all located in a 500m radius and revolve around the user's work location; clusters 5 to 7 appear in the early morning and are also close together, which suggests they pertain to the neighborhood where the user lives.

Unless the points in an invisible cluster can be individually matched to tagged Foursquare check-ins, the underlying venue of an invisible cluster is not apparent from the raw data. We attempt to automatically attach a label to invisible clusters by applying a simple time-of-day-based heuristic, similar to the one suggested in [2], in order to identify clusters that appear in places such as "home" or "work". To do so, we first aggregate the coverage of all clusters that insist on the same location. Our rationale is that aggregates of invisible clusters with significant coverage might represent important locations where users don't usually perform check-ins. We define simple rules that classify invisible clusters into four groups:

- Long (aggregate duration > 3h) daytime (7AM-6PM): WORK group
- Long (aggr. duration > 3h) nighttime, or longer than 10 hours: HOME group
- Short (aggr. duration < 3h) daytime (7AM-6PM): other daytime feature
- Short (aggr. duration < 3h) nighttime: other nighttime feature

Figure 5 presents a breakdown of the features detected among all the users with clusters in their GTW traces. We observe that 80% of the users in the WD and 90% in the WE dataset show at least one home cluster (only 11 of the 155 WD users have more than one). Work clusters by the above definition appear in about 8% of the users in the WD dataset and just for one user in the WE dataset. Shorter

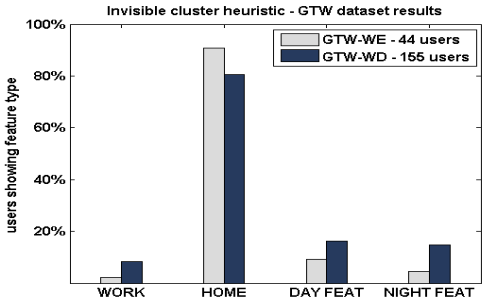

Figure 5. Invisible clusters: results of work-home heuristic on GTW dataset

daytime features that do not get classified as work are present in respectively 16% and 9% of the users while nighttime features appear in 15% and 5%. As we lack a ground truth for validating the outcomes of the heuristic, we argue nonetheless that results about home clusters are plausible given the amount of events per user that are collected in our GTW traces.

B. Tagging Foursquare Clusters with Activities

To present the observed user activities from the routines of our 825 users, we plot the their breakdown over the course of a day (Figure 6). In absolute terms, we observe that the maximum amount of routine clusters is observed in the early morning, with 229 unique users showing at least one activity cluster between 6AM and 8AM and 200 between 8AM and 10AM. A second peak, much smaller, can be seen in the early afternoon. The number of observed clusters drops sharply during the evening, hits a minimum between midnight and 2AM, and picks up only before dawn (4AM to 6AM). We remark that the types of activities we detect mostly cover the daytime and primarily capture salient features of working life, such as traveling to work in the morning, breakfast, lunch, and traveling back in the evening. Our results are consistent with the known Foursquare check-in trends, and confirm the dearth of home and work check-ins. Comparing against published results about observed Foursquare check-in venue types during the day [11] we notice that popular but less regular activities, such as dining out and nightlife, are indeed less represented in the global routine breakdown, as one might expect.

V. Predicting User Behavior

In the previous sections, we described a method to extract clusters of events that capture some salient features of social

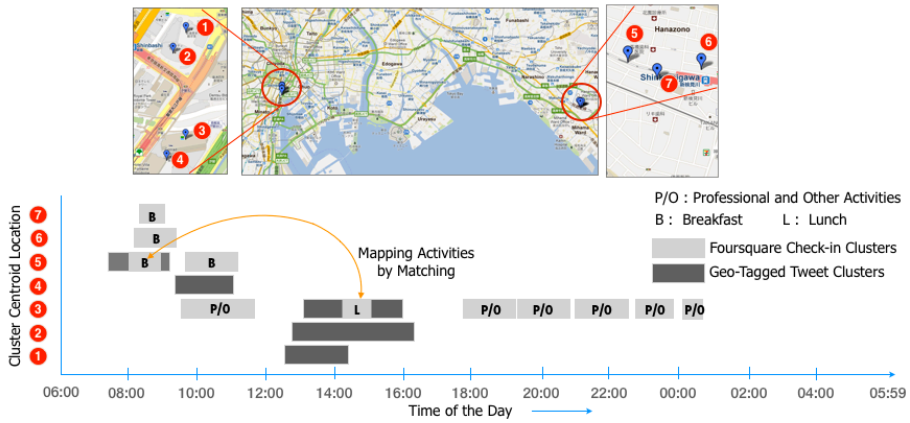

Figure 4. Example of user routine extracted by differential cluster analysis of 4SQ and GTW traces. Clusters from the 4SQ dataset are identified by the associated activity, as shown in the legend. Non-matched darker clusters are "invisible" in the Foursquare check-in trace and emerge from the GTW dataset

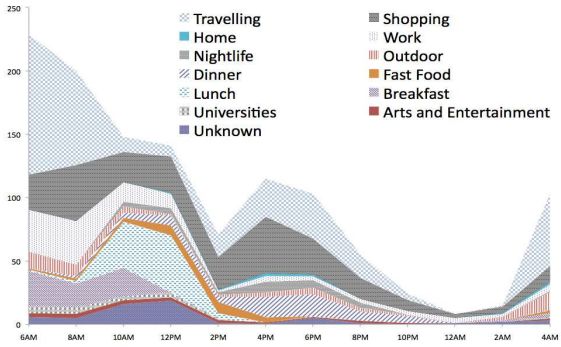

Figure 6. Breakdown of global daily routine features in the 4SQ-ALL trace

networking traces, characterizing the temporal incidence of routine activities in the users' daily lives. We now evaluate the suitability of routines for predicting future user behavior.

A. Predicting User Behavior from Routines

Routines identify steady features of a user's daily life. Predictions can thus be obtained in our model simply by extending the descriptive value of the clusters that have been observed on a set of **training data** to be prescriptive about user activity in an independent corpus of **testing data**. We define a prediction as an *outstanding expectation about a certain feature of user behavior*. As any feature cluster $C_X(t_{start}, t_{end})$ exists in the time domain, a prediction based on the cluster associates the expected value of the property X of an event to the contiguous range of time values, between t_{start} and t_{end}, that are covered by the cluster. A predicted outcome is then considered successful if, at the time t_p when a given event is observed in the testing data, the value of the event property X matches the value of C_X for *at least one* of the clusters that are active at t_p.

Our reliance on discrete clusters as the basis of predictions means that there will be time values for which there is no expectation. The main limitation of our approach is that we will not attempt guesses outside the boundaries of the observed clusters. As a consequence, we account for these missing predictions with a *hit rate* metric, which tracks the fraction of guesses attempted over the total number of possible guesses. In the evaluation of the *success rate* of predictions, we will therefore only consider the cases in which a guess is made (*i.e.* a hit) which turns out to be correct. As a further measure of the ability to provide predictions for a given user, we introduce a *coverage* metric that describes the ratio between the union of all clusters' temporal support and a 24-hour period, which corresponds to the fraction of the time for which a prediction is available.

B. Prediction Results

As we discussed previously In order to users displaying significant features in their daily behavior, we will focus on the set of users who belong to the STV category, obtained from the global dataset using the segmentation methods described in Section II-C, as our sample for the following evaluation. We compute the prediction results for the Foursquare check-in dataset (4SQ) in the weekdays (WD), weekends (WE), and WD-WE combined (ALL) traces. Due to the moderate richness in events of the STV users, we expect them to be a representative example of a typical mode of interaction by regular users of social networking applications.

Figure 7 shows the results of spatial and venue category forecasts based on the 4SQ-WD dataset in terms of hit and success rate for a training period length M, ranging from one to six months. The value of k used in this case is five, and all reported results are averages of experiments conducted with $\frac{12}{M}$-way randomization in the choice of non-overlapping blocks of M contiguous months of training data. We can observe from the picture how both hit and success rate increase for increasing values of M: after six months of training, the hit rate of category predictions is above 80%, with a success rate of more than 66%. In the case of spatial predictions, both hit rate and accuracy are lower: for the same six-month training

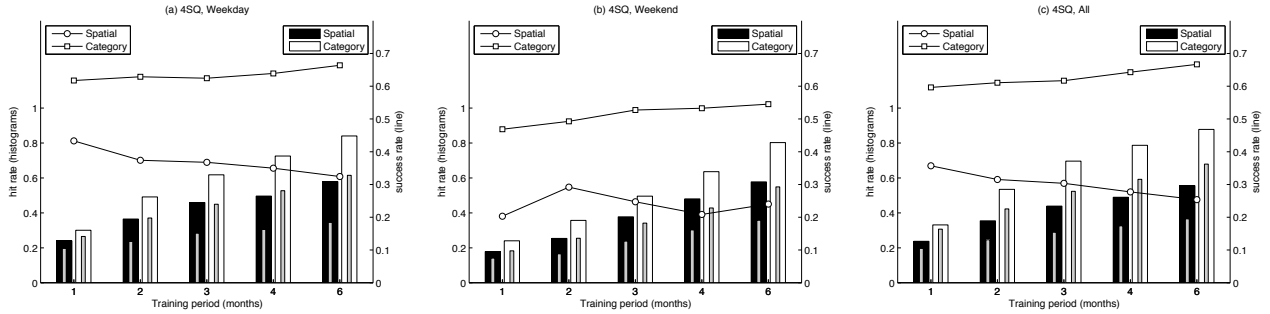

Figure 7. Category and Spatial predictions: hit and success rate results for the 4SQ dataset (WD, WE, ALL) - coverage metric in grey

length, a valid prediction exist only for 58% of the events, with a success rate of about 32%. Finally, we notice that the average coverage rate of the predictions, represented in Figure 7 by the narrow grey bars, increases with the length of the training period but is always strictly smaller than the average hit rate: this result confirms that check-in events are unevenly distributed throughout the day, and that predictive clusters tend to cover periods in which check-ins are more likely.

C. Adaptive Spatial Predictions

Observing the outcomes of spatial and category predictions, we noticed that some activities lend themselves reliably to Spatial predictions, while others show little consistency in the prediction outcome: for instance, "Travel Spots" such as train and subway stations are a remarkably stable feature of user routines; on the other hand, "Food" venues sometimes do not allow good spatial predictions, as for instance an epicurean user could be sampling every time a different restaurant from a large list of favorites. We now investigate the correlation between our ability to formulate both category and spatial predictions in order to improve the reliability of the latter.

We introduce a second step in the training process of our clusters: after the two sets of category and spatial clusters have been formed, we use the training data for a second time to estimate the probability that an event for which a category prediction exists will lend itself to a correct spatial prediction on the training set. For each user, we compile a table that assigns every category cluster a *reliability score* for spatial predictions, defined as the fraction of correct guesses over the relevant events in the training set. After training, the value of the reliability score for of a category cluster will inform the decision of whether or not a spatial prediction should be attempted, thus trading a lower hit rate against an increase of the success rate of the fewer spatial predictions attempted. In Figure 8 we compare the results in terms of hit and success rate for normal and adaptive spatial predictions made over the 4SQ dataset with a reliability threshold of 50%. We can observe in most cases the expected slight reduction in the hit rate, which is compensated by a noticeable increase in the successful prediction rate.

VI. Related Work

The participatory approach to social data collection has rapidly become popular in the literature, producing a wealth of related studies. In [13], Li and Chen quantitatively analyze large-scale LBSN traces and offer a generic perspective on user profiles, update activities, mobility characteristics, social graphs, and attribute correlations. Eagle [14] shows ways to use social sensing to study human behavior, namely to discover daily activity patterns, to infer relationships, and to determine significant locations. Ye *et al.* [15] presents a semantic annotation technique for location-based social networks, where user check-in behavior is analyzed to extract the significance of individual places and the implicit relationships among similar places. In [16], Ferrari *et al.* study 13 million geo-tweets collected from New York Metropolitan area to understand urban crowd behavior. Their elegant probabilistic topic-model analysis uncovers patterns that highlight hidden urban dynamics and recurrent spatio-temporal crowd phenomena in a urban scenario. Joseph *et al.* [17] apply Latent Dirichlet Analysis (LDA) on Foursquare data from check-ins in New York City and the Bay Area to produce latent collections of people with similar interests and social activity trajectories. In [11], Noulas *et al.* investigate user check-in dynamics in an attempt to uncover meaningful spatial-temporal patterns of user mobility in urban spaces. While their work focuses on understanding *collective* location dynamics, our approach is tailored to analyzing individual user's recurring activity patterns. The work with the closest goal to ours, although based on a different type of data source, is [18]. Its authors used LDA to discover activity routines in the Reality Mining data set from MIT, a dataset containing one year worth of mobile phone traces from 94 volunteers.

Literature is also rich on the subject of predicting future locations and activities of users based on traces collected opportunistically from either mobile or social networks. In [19], Ying *et al.* propose an approach for predicting next location of an individual based on both geographic and semantic features of her trajectories. The predicting technique leverages clusters of similar users to predict a user's next location. In [8], Scellato *et al.* introduce NextPlace, a spatial-temporal framework based on non-linear timeseries analysis of start times and duration of

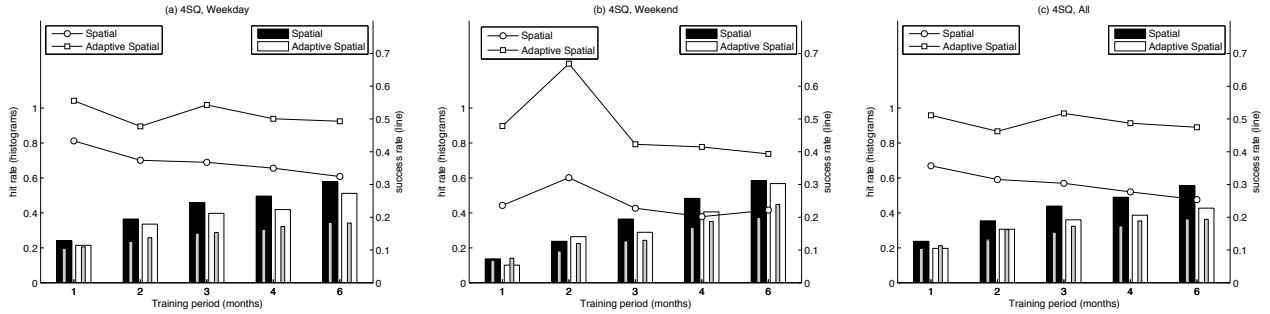

Figure 8. Adaptive Spatial predictions: coverage, hit, and success rate results for the 4SQ dataset (WD, WE, ALL) - coverage metric in grey

visits to significant locations. Their technique can forecast the next location of a user, as well as the length of time the user will spend there. Also relevant is [20], where authors examine GPS traces to predict a driver's destination, based on her habits and general driving behavior. Although accurate, these studies require sources of location data with much higher spatial and temporal resolution than can be derived from social network traces. Participatory sensing approaches, such as the one we presented in this paper, operate instead on datasets composed of relatively sparse and coarse-grained location samples.

Presently, our method does not rely on a global trajectory pattern base, nor does it consider the interplay between different users across the same locations. We are considering the use of homophily-based learning, where similar users' collective check-in histories can be leveraged to recognize individual user activity. We expect techniques such as presented in [21] to be effective at increasing the inference density.

VII. Conclusion

Social networks have a huge potential as sources of insight on human behavior and recurring activity patterns. A participatory approach to data collection applied to social networks, a passive and non-invasive method, has the ability to easily scale up and cover large human populations. In this paper we performed an analysis of significant spatial and temporal features emerging from a year-long trace from 825 users of the Foursquare and Geo-tagged Twitter social networks. We presented a method to extract user routines and identify the activities associated with their spatial-temporal features. Finally, we explored ways to extend the benefits of available metadata across different traces by matching corresponding points and features that appear in different datasets.

References

[1] D. Ashbrook and T. Starner, "Using GPS to learn significant locations and predict movement across multiple users," *Personal and Ubiquitous Computing*, vol. 7, no. 5, pp. 275–286, Jan 2003.

[2] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying important places in peopleś lives from cellular network data," in *Pervasive Computing*, 2011, vol. 6696, pp. 133–151.

[3] J. H. Kang, W. Welbourne, B. Stewart, and G. Borriello, "Extracting places from traces of locations," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 9, no. 3, pp. 58–68, Jul. 2005.

[4] M. Kim, D. Kotz, and S. Kim, "Extracting a mobility model from real user traces," in *Proc. INFOCOM 2006*, april 2006, pp. 1–13.

[5] L. Liao, D. Patterson, D. Fox, and H. Kautz, "Building personal maps from GPS data," *Annals of the New York Academy of Sciences*, vol. 1093, pp. 249–265, 2006.

[6] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti, "Wherenext: a location predictor on trajectory pattern mining," in *Proc. of KDD-09*, New York, NY, USA, 2009, pp. 637–646.

[7] S. Phithakkitnukoon, T. Horanont, G. Lorenzo, R. Shibasaki, and C. Ratti, "Activity-aware map: Identifying human daily activity pattern using mobile phone data," in *Human Behavior Understanding*, 2010, vol. 6219, pp. 14–25.

[8] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. T. Campbell, "Nextplace: a spatio-temporal prediction framework for pervasive systems," in *Pervasive computing*, 2011, pp. 152–169.

[9] N. Bila, J. Cao, R. Dinoff, T. K. Ho, R. Hull, B. Kumar, and P. Santos, "Mobile user profile acquisition through network observables and explicit user queries," in *MDM*, 2008, pp. 98–107.

[10] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *In Proc. of KDD-96*, pp. 226–231, Jan 1996.

[11] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, "An empirical study of geographic user activity patterns in foursquare," in *Proc. of AAAI ICWSM 2011*, 2011, pp. 570–573.

[12] J. Lindqvist, J. Cranshaw, J. Wiese, J. Hong, and J. Zimmerman, "I'm the mayor of my house: examining why people use foursquare - a social-driven location sharing application," in *Proc. of Conference on Human factors in computing systems (CHI 2011)*, 2011, pp. 2409–2418.

[13] N. Li and G. Chen, "Analysis of a Location-Based Social Network," in *2009 International Conference on Computational Science and Engineering*, Aug. 2009, pp. 263–270.

[14] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems," *Personal Ubiquitous Computing*, vol. 10, pp. 255–268, 2006.

[15] M. Ye, D. Shou, W. Lee, P. Yin, and K. Janowicz, "On the semantic annotation of places in location-based social networks," in *Proc. of KDD-11*, 2011, pp. 520–528.

[16] L. Ferrari, A. Rosi, M. Mamei, and F. Zambonelli, "Extracting urban patterns from location-based social networks," in *Proc. of ACM SIGSPATIAL LBSN '11*, New York, New York, USA, Nov. 2011, pp. 9–16.

[17] K. Joseph, C. H. Tan, and K. M. Carley, "Beyond "local", "categories" and "friends" : clustering foursquare users with latent "topics"," in *Fourteenth International Conference on Ubiquitous Computing (Ubicomp 2012)*, New York, NY, USA, Sep. 2012, p. 919.

[18] K. Farrahi and D. Gatica-Perez, "Discovering routines from large-scale human locations using probabilistic topic models," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 1, pp. 1–27, Jan. 2011.

[19] J. J.-C. Ying, W.-C. Lee, T.-C. Weng, and V. S. Tseng, "Semantic trajectory mining for location prediction," in *Proc. of ACM SIGSPATIAL GIS '11*, New York, NY, USA, Nov. 2011, p. 34.

[20] J. Krumm and E. Horvitz, "Predestination: Inferring destinations from partial trajectories," *In Eighth International Conference on Ubiquitous Computing (UbiComp 2006)*, Jan 2006.

[21] D. Lian and X. Xie, "Collaborative activity recognition via check-in history," in *Proc. of ACM SIGSPATIAL LBSN '11*, 2011, pp. 45–48.