Inferring Human Mobility Patterns from Anonymized Mobile Communication Usage

Yuzuru Tanahashi
ViDi Research Group
University of California Davis
ytanahashi@ucdavis.edu

James R. Rowland
AT & T Labs – Research
Florham Park, New Jersey
jrr@research.att.com

Stephen North
AT & T Labs – Research
Florham Park, New Jersey
north@research.att.com

Kwan-Liu Ma
ViDi Research Group
University of California Davis
ma@cs.ucdavis.edu

ABSTRACT

Anonymized Call Detail Records (CDRs) contain positional information of large populations and therefore have been extensively analyzed to understand human mobility. Due to the temporally sparse and spatially coarse nature of the data, most of these studies have focused on primitive aspects of movements such as travel distance and speed. Incorporating underlying geographic information in these analyses would allow analysts to put these movements into context and to gain deeper insight into how metropolitan areas function. In this paper, we present a set of procedures for inferring mobile users' mobility patterns while retaining the context of underlying geography. We apply these methods to our case study on New York City anonymized CDRs. We find that our methods verify current areal semantics and commuting rush-hour patterns, and we also derive further implications regarding geographic, demographic, and other effects on human mobility.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Algorithms, Theory

Keywords

Mobile computing, call detail records, human mobility

1. INTRODUCTION

Location data generated by mobile communication services has enabled researchers to investigate patterns of human movement within a metropolitan area. A better understanding of how, where, and when people move on a daily basis, especially in densely populated areas, could lead to improvements in urban planning, transportation infrastructure design, and assessments of environmental impact.

Traditionally, human mobility studies have been carried out with surveys. More recently, the data collection process has been automated by various location-aware mobile devices. For instance, there are many studies on human mobility which analyze personal location data collected from global positioning system (GPS) devices [1, 11, 12]. Although location-aware mobile devices are increasingly common, and more people have begun to share location information on the web, revealing personal information raises privacy concerns [18, 21]. Therefore, datasets based on such shared records are potentially biased not only toward the people who choose to reveal their location, but also toward the places they choose to reveal.

Anonymized Call Detail Records (CDRs) from calls and SMS messages provide a rich source of information regarding locations of mobile phones. This data is generated as a part of the infrastructure of mobile communication networks and has a well-developed data management framework to support billing, fraud avoidance, and network engineering. Since CDRs are recorded by all mobile communication carriers, this data is also ubiquitous not only among different cities but also at different time periods where mobile communications are available. Therefore, extensive research has been aimed at methods for analyzing anonymized CDRs.

Due to the assumed physical proximity of a mobile phone and its user, many researchers have applied anonymized CDRs as a proxy for a mobile user's locations (see section 2). However, the analysis of anonymized CDRs with respect to human mobility involves several challenges. These challenges are mainly oriented towards two properties that are unique to anonymized CDRs. One property is that anonymized CDRs only contain the location of the cell tower which is handling the voice call or SMS event. Therefore, anonymized CDRs can only provide an approximate location of the mobile user based on the assumption that a mobile user is near the cell tower. The other property is that anonymized CDRs are only recorded when a communication event takes place. Thus, there is no location information available during the time the phone is not in use.

In this paper, we present a set of methods for extract-

MoMM2012, 3-5 December, 2012, Bali, Indonesia

ing human mobility patterns from anonymized CDRs and we demonstrate the application of these methods through a case study using real-world data. The methods we introduce in this paper include a new algorithm for partitioning maps into spatially contiguous regions, a scalable model for inferring mobile users' aggregate daily movement routines (*life-patterns*), and a metric for quantifying the predictability of a mobile user's location. In the case study, we apply visualization techniques to the extracted information and compare the represented information with patterns already known by domain experts.

The dataset we use in this study consists of real-world anonymized CDRs from approximately 120,000 random mobile users whose billing addresses belong to the New York metropolitan area. This dataset excludes phones registered to businesses to avoid situations where usage will not give an accurate view of overall mobility patterns. The anonymized CDRs are based on records from March 15 to May 15, 2009, a sixty-two day period.

2. RELATED WORK

While human mobility analysis using specialized spatio-temporal datasets, such as GPS logs, has shown significant potential in conveying human mobility patterns, due to the overhead of collecting such detailed datasets, many researchers have experimented with applying other data sources as a proxy for human mobility. Brockmann *et al.* described a quantitative study on human migration in the United States by analyzing the circulation of bank notes [8]. In a more recent approach, by analyzing human mobility data extracted from collected logs of microbloggers using location aware smartphones, Fujisaka *et al.* observed gathering and dispersing movement patterns frequently occurring in urban areas [12].

Due to its ubiquitous existance and its ability to present information at various spatial scales, there is interest in studying anonymized CDRs to help understand aggregate human movement. Many researchers have conducted various statistical analyses on large collections of anonymized CDRs to better model human mobility. González *et al.* discovered, in contrast to the pervasive mathematical models, individual travel patterns fall into a single spatial probability distribution and humans follow simple patterns for daily travel paths indicating individual daily travel patterns are feasible [14]. Song *et al.* conducted an analysis on the predictability of human mobility using anonymized CDRs and concluded that although the predictability of an individual's movements varies depending on various factors (e.g., the frequency of phone calls and the range of movement) there is regularity in daily movement habits [27, 26]. By comparing anonymized CDRs from different cities, Isaacman *et al.* found significant differences in human mobility habits between Los Angelenos and New Yorkers [16].

While many studies on large collections of anonymized CDRs have revealed various aspects of human mobility patterns, most of these studies focused on the spatial derivatives of human mobility and without taking into account the underlying geography. Since geographic features have significant effects on how and where people move, geographic information is essential for analysts to investigate contextual human mobility trends in urban areas. In one recent attempt to analyze human mobility from anonymized CDRs, Martino *et al.* introduced a visualization tool which allows analysts to explore partial trajectories of individual mobile users in combination with aggregated statistical information such as the approximated population distribution [19]. However, the trajectories in their studies were limited to those of mobile users with exceptionally heavy usage.

In other attempts to analyze anonymized CDRs, some researchers have also looked into the possibility of utilizing the cell phone usage information from anonymized CDRs to characterize human activities in urban areas. Reades *et al.* developed a method for analyzing regional characteristics based on weekly phone activities and discovered that different areas have different cell phone activity trends [24]. Girardin *et al.* demonstrated how aggregated anonymized CDRs can still reveal the differences in points of interest between local residents and tourists in New York City [13]. Becker *et al.* proposed a visualization for depicting the daily anonymized CDRs for each cell tower and showed how cell towers pointed toward different areas (downtown and high school) have different daily phone activity trends [5].

This paper focuses on presenting the methods for processing a collection of anonymized CDRs to infer mobility patterns with respect to the underlying geography. In our study, we also include anonymized CDRs of mobile users with low usage and discuss the correlation between usage and location predictability in our model. This work contributes to the field by providing a means of analyzing anonymized CDRs with respect to human mobility patterns in urban areas while retaining geographic context of the movements. We demonstrate our approach on real-world data.

3. DATA

Anonymized CDRs are logs of mobile phone activities collected mainly for billing purposes. Each anonymized CDR consists of six items regarding the event: user id, date, time, duration of the event, and the locations of the cell towers which handle the starting and the ending of the event. Table 1 shows an example of this data.

Table 1: Anonymized Call Detail Records

ID	Date	Time	Duration (min)	Location	
				Start	End
1	3/16	21:49	10	37.52,-80.87	37.54,-80.90
4	3/16	22:17	30	37.22,-80.85	37.34,-80.90
1	3/16	23.39	0	37.14,-80.64	37.14,-80.64
2	3/17	00:12	13	37.22,-80.85	37.22,-80.85

The data used in this study includes approximately 45 million anonymized CDRs based on about 120,000 mobile users during the sixty-two day period between March 15 and May 15, 2009. These mobile users were randomly selected from non-business mobile phones, whose billing addresses are in the vicinity of New York City. The subject area consists of 187 zip codes in New York and New Jersey, and covers approximately 660 square miles.

Although the average number of daily phone events per mobile user is six, the actual distribution of daily phone usage has a long tail and the majority of mobile users have less daily mobile activities than the average. Hence, most mobile users make fewer than five phone calls while some heavy users make more than 10 phone calls per day. Figure 1 shows the histogram of the average daily phone usage.

This sparsity in the data makes the task of inferring mobile patterns more challenging.

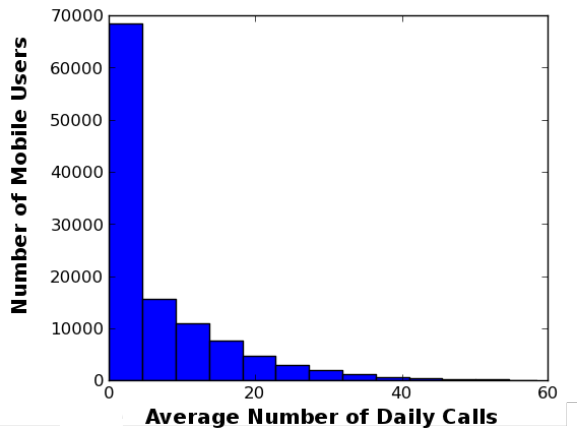

Figure 1: A histogram of mobile subjects average usage. Most users make fewer than five phone calls per day.

Due to the privacy concerns inherent in personal positional data [3, 31], in order to further enhance privacy and strengthen anonymity, the processing of anonymized CDRs involves multiple layers of data abstraction. In this research, we enhance anonymity through three types of data abstraction:

- spatial abstraction,
- temporal abstraction, and
- personal abstraction.

Spatial abstraction is performed by reducing the spatial resolution of the data (see section 4.1). By converting the locations in anonymized CDRs into symbolic regions that cover large areas, we obscure identifiable locations in the data. Temporal abstraction is handled by removing dates from the anonymized CDRs and applying a diffusive function to the observations in the dataset over time (see section 4.2). By obscuring the timestamps in anonymized CDRs, we can reduce distinctive features in phone call habits which may disclose user identity. Personal abstraction is applied by aggregating the data into a statistical dataset which summarizes the information of a collective population. This agglomeration inhibits extraction of the underlying personal information. We conduct our case study in section 5 based on this aggregate data. Although we discuss these abstraction methods in section 4 focusing on their application to anonymized CDRs, we believe these methods can also be applied to other positional data, such as GPS logs, to increase anonymity and improve privacy.

4. METHODOLOGIES

In this section we describe the methods we developed for processing anonymized CDRs to extract spatio-temporal patterns. In this study, the anonymized CDRs go through three processing steps.

The first step is the conversion of location information. In this step each cell tower location in the anonymized CDRs is translated into a symbolic region. Converting the location information in anonymized CDRs to a symbolic region has several benefits [29]. In this study, abstracting the location information by lowering spatial resolution can not only enhance anonymity, as discussed in section 3, but also reduces the effects of spatial uncertainty inherent in anonymized CDRs.

In this research, we propose a new method for partitioning maps into modular spatial units based on human movement. This computation allows the symbolic regions to maintain contextual features of the underlying geography.

The second step is the modeling of mobile users' daily trajectories. For this step, we propose a probabilistic model for approximating daily mobility patterns (*life-patterns*). A mobile user's trajectory over time is simulated by its probabilistic *life-pattern*. By putting the focus on mobile users' daily routines instead of actual movements, our method augments the scarcity of observations by aggregating information in anonymized CDRs from different days.

The final step is filtering. We eliminate subjects with less consistency in their daily routines. This allows us to focus on the mobility patterns of subjects whose patterns are more consistent. In this step, we apply entropy as a metric to quantify the consistency and the predictability of each mobile user's derived *life-pattern*.

Although entropy-based analysis on predictability in human mobility from anonymized CDRs has been discussed previously [27], the dataset used in this previous research only included subjects whose average number of daily phone calls was at least 12. As we can be see in figure 1, mobile users with more than 12 usage events per day are a small fraction of the actual population of mobile owners and can be considered extreme heavy users. In our study, we include mobile users with infrequent phone calls to avoid selective sampling. In addition to this fundamental difference in data sampling, we also use a different spatial resolution for projecting human mobility patterns and propose a prediction model that combines observations from different days. Therefore, the predictibility results from our study do not agree closely with the results shown in Song *et al.* [27]. Their findings are that most people have consistent daily travel patterns; this supports our approach of extracting daily mobility patterns rather than actual movements.

4.1 Clustering Zip Codes

One of the primary challenges in geographic data analysis is the appropriate choice of spatial resolution. While contextual partitioning can be accomplished using common clustering methods (e.g., [2, 24]), few methods guarantee spatial contiguity of clusters. Such contiguity is critical for analyzing human movement to capture both spatial and geographic context of movements. For example, to assist the analysis of the geographic interaction patterns of human movement and pandemic spreads, Guo proposed a clustering method for partitioning maps which maintains each region's spatial contiguity [15].

We introduce a new method for clustering locations into spatially contiguous regions. Our method contextually segments maps into community areas by extracting the underlying structures based on regional human mobility. By partitioning maps into modular regions, analysts can effectively shift their focus on the movements relevant to the area of interest without being overwhelmed by the high-frequency local movements in the data. For instance, in analyzing human mobility with respect to the whole country of the

U.S., analysts would prefer to focus on inter-state migration patterns rather than migration patterns between neighborhoods, since the latter would mostly consist of insignificant migrations and would bury important information to the analysis in the overcrowding data.

Our method is based on a graph partitioning algorithm which focuses on optimizing the overall modularity of the clustering result. Modularity is a widely used metric proposed by Newman and Girvan for evaluating graph clustering [23]. Equation 1 shows how the modularity of a clustering result C can be calculated given a graph $G(V, E)$, where V represents the vertices and E represents the edges. Here, $E(c)$ is the number of the edges that are within the cluster c, and $E(c, c')$ is the number of the edges that connect clusters c and c'. A detailed overview of modularity clustering is described by Fortunato [10]. Ideally, maximizing modularity partitions complex networks into natural communities. However, finding an optimal partitioning with the highest modularity is NP-complete [7]. Therefore, many approximation algorithms for modularity clustering have been proposed. In this study, we apply two well-developed approximation algorithms, the *Greedy Technique* [22] and *Extremal Optimization* [6].

$$\sum_{c \in C} \left[\frac{|E(c)|}{|E|} - \left(\frac{|E(c)| + \sum_{c' \in C} |E(c, c')|}{2\,|E|} \right)^2 \right] \quad (1)$$

The two graphs used in our algorithm are the *activity graph* and the *space graph*. The *activity graph* represents human mobility between different areas. The *space graph* represents the spatial contiguity between different areas. Both graphs share the same vertices, which are zip code[1]areas. However, the edges in the *activity graph* represent human movement between zip codes, and edges in the *space graph* represent zip code adjacency. In the *activity graph*, edges are also weighted by the volume of traffic.

Our algorithm is iterative and consists of three steps. The steps are:

1. Initial partitioning
2. Smoothing
3. Refining

In the initial partitioning step, we apply modularity clustering to the *activity graph*. The clusters represent regions with high intra-region traffic. However, these areas are not guaranteed to be spatially contiguous.

The smoothing step transforms the spatially discrete regional clusters into a series of spatially contiguous regions. This step starts by translating the clustering results into the *space graph*. Next, each cluster's core node is determined. In our case, we calculate the spatial center of mass of each clustered region, and choose the vertex closest to it. Finally, each vertex is tested for spatial contiguity with its cluster's core node by applying a breadth-first search on all vertices within the same cluster. If the vertex is spatially discontiguous with its cluster's core node, the vertex is assigned an alternative cluster from the vertices adjacent to it in the *space graph*. For selecting the cluster from potential candidates, we calculate the overall modularity for all possible exchanges and select the one that maximizes the modularity in the *activity graph*. This step is an iterative procedure, and continues until all vertices are assigned to spatially contiguous clusters.

The refining step is similar to the Kernighan-Lin heuristic [17] and uses the spatially contiguous clusters derived in the previous step. In this step, each vertex is tested to see if moving it to another cluster will improve the overall modularity of the *activity graph*. By limiting the candidate clusters for to adjacent vertices in *space graph*, this step maintains the overall spatial contiguity of each cluster in the *activity graph*. This step is also an iterative procedure, and continues until the *activity graph* is no longer improved.

Table 2: Modularity Yield

Approximation Technique	Modularity
Greedy Technique	0.6526
Extremal Optimization	0.6570

Table 2 shows the overall modularity of the partitioning results of the map based on the two methods *Greedy Technique* and *Extremal Optimization*. Although *Extremal Optimization* has a slightly higher modularity, the difference is very small.

Figure 2 shows three maps which are partitioned by geographic county borders, *Greedy Technique*, and *Extremal Optimization*. These maps cover six counties: Bergen, Hudson, New York (Manhattan), Bronx, Queens, and Kings County (Brooklyn). We also assign a name to each area in the seven regions derived by *Extremal Optimization* based on the area they cover. The names of these regions are Upper New Jersey, Lower New Jersey, Uptown Manhattan, Downtown Manhattan, Bronx, Queens, and Brooklyn. These names are also shown on the map in figure 2.

Both maps partitioned by our clustering method in figure 2 significantly resemble the partitioning of the actual counties. Although there are some evident geographic features which separate these communities (e.g., the Hudson River) some community borders do not correspond to geographic features. For example, in both clustering results, Downtown and Uptown Manhattan are clearly separated despite their spatial contiguity. Queens and Brooklyn also form separate communities in both clustering results, although they have no significant geographic features that separate them.

The similarity between our clustering results and the actual county borders has some interesting implications. Since our clustering is based on modularity in human movement, this alignment indicates that most people move within county borders. While this may be simply because most people prefer to travel less and stay within their community, it may also be due to a reason more social, such as transportation infrastructure making certain areas more accessible than others. For example, the metro stations significantly enhance human mobility along railways. Many public buses also have fixed routes within counties, which inherently limit connections with other counties.

We have also discovered that although the border between Uptown Manhattan and the Bronx in our *Extremal Optimization* partitioning does not align precisely with the ac-

[2]In this research, we translate each cell tower to its located zip code based on the collaborating analysts' request. The traffic between zip codes is the summation of mobile users' subsequent anonymized CDRs containing corresponding cell towers.

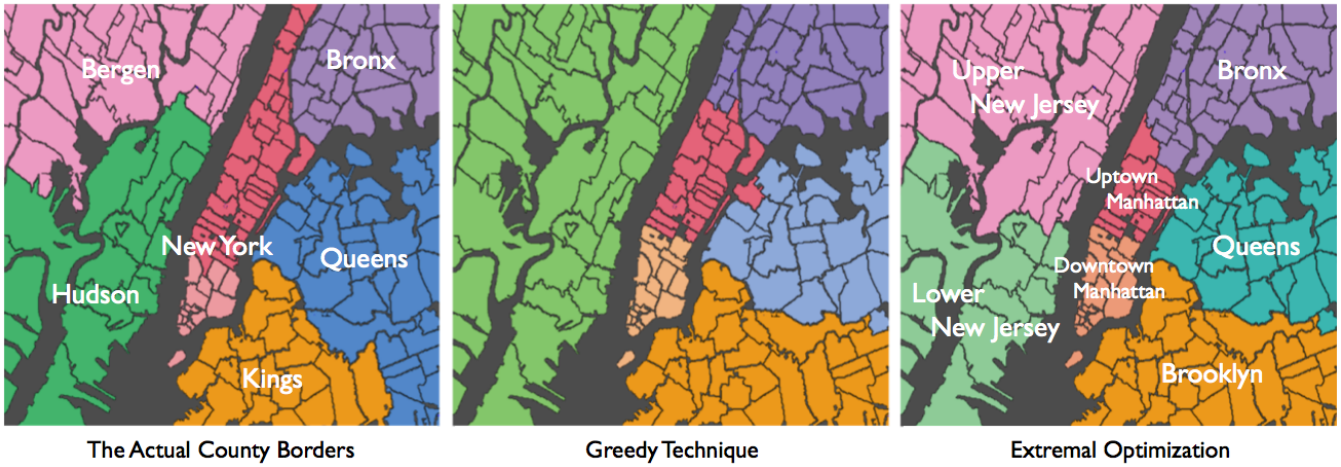

Figure 2: Partitioning results of the map based on the actual county borders, *Greedy Technique*, and *Extremal Optimization*. Each polygon in the map represents a unique zip code area and is color-coded based on which community it belongs.

tual county border, which is along the river, the border of our clustering result aligns with the border in the demographic mapping based on the Census Bureau's American Community Survey from 2005 to 2009 [28]. A visual representation of this demographic distribution can be seen in The New York Times "Mapping America" [20]. Based on this survey, the major ethinicity whose residents are north of the border (Bronx) are African-Americans and Hispanics, while the south of the border (Uptown Manhattan) are mostly occupied by Caucasians. This implies that human mobility is not only affected by geographical features (e.g., rivers and mountains), but also by demographical features.

4.2 Inferring Life-patterns

One of the difficulties in translating anonymized CDRs into movement data is the irregularity of mobile users' daily phone usage activities. Different people make phone calls at different times, and usage by may even differ from day to day for individuals. Therefore, it is practically impossible to infer people's movements based on the fragments of positional information observed in daily anonymized CDRs.

However, most people have largely consistent daily mobility routines [27]. For example, commuting patterns typically do not change drastically from one day to another. Such highly structured and consistent daily routines will be referred to as *life-patterns*. In this research, we have developed a method for inferring *life-patterns* by combining multiple observations of usage over different days. In this section we describe the model we use for translating a collection of personal anonymized CDRs into probabilistic approximations of *life-patterns*.

This method applies a Naïve Bayes Model to calculate the probability of an anonymous subject's location at a given time. This method is not only simple and scalable, but also suitable for aggregated data with highly irregular records. We choose this model over common models for predicting movements, such as Hidden Markov Models (HMMs) that focus on modeling movements based on correlations between consecutive observations in the data. While this method has proven effective for most movement data, since consecutive positions observed in anonymized CDRs often have a significant time gap[2], correlations between these observations often are irrelevant to the actual movement of mobile users.

In this model, each mobile user's *life-pattern* is represented as a series of probabilistic spectra which construct a matrix M. M consists of $|L|$ rows and $|T|$ columns, where L is the set of symbolic regions and T is the set of time segments. In this matrix, the value of $M(l,t)$ is the probability of the user being at location l at time t.

In our case study, L consists of seven regions, and T consists of 144 time segments. The seven regions are based on the communal regions derived from our map partitioning algorithm described in section 4.1 using *Extremal Optimization*, Upper New Jersey, Lower New Jersey, Bronx, Uptown Manhattan, Downtown Manhattan, Queens, and Brooklyn. Each time segment in T is 10 minutes; 144 segments collectively represent a whole day. We chose this segmentation based on the frequency of the New York City subway traffic. The consecutive columns in matrix M also correspond to consecutive time segments of the day.

The generation of a particular *life-pattern* matrix involves two steps. The first step is the translation of anonymized CDRs into the corresponding matrix entries. The second step is the smoothing of the probabilistic distribution over the time axis. This step not only allows our method to approximate the *life-pattern* probability distribution for the time segments with few observations, but also increases the data anonymity regarding the time stamp information as discussed in section 3. Also, at end of each step the probabilities in each column are normalized among all locations.

Equation 2 shows the first step in our model for calculating the probability of a mobile user being at region l at time t. This equation includes the application of Laplace correction to all probabilities. Here, the function C returns the number of the mobile user's anonymized CDRs which match the condition described in the input argument.

[2]Most consecutive pairs of anonymized CDRs have more than 4 hours between them.

$$M(l,t) = \frac{C(L = l \cap T = t) + 1}{C(T = t) + |L|} \quad (2)$$

Although this translation of anonymized CDRs to a probabilistic *life-pattern* matrix conveys many mobile users' temporal dynamics of their location likelihoods, several issues due to the irregular observations from anonymized CDRs remain in the calculated *life-pattern* matrix. Here, we refer to the sequence of regions with the highest likelihood over each column in the *life-pattern* matrix as the mobile user's most likely *life-pattern*. One of the problems is that the direct translation produces a *life-pattern* with frequent fluctuation in the probabilistic spectra, which often result in an unrealistic frequent switching of regions as its most likely *life-pattern*. Another problem is, due to the diversity in usage habits among mobile users, many mobile users who actually share similar *life-patterns* can still be as far apart in Euclidean space as a user with a completely different *life-pattern* because their time segments with likelihood peaks do not align. This can be problematic when analysts want to cluster mobile users based on their *life-patterns* to conduct a contextual filtering for analysis.

To address these problems, we apply a kernel smoother to the probabilistic *life-pattern* matrix in the second step of our model. Kernel smoothing is a conventional and effective technique for estimating missing values in data based on noisy observations [30]. In our method, kernel smoothing allows an observation made at a certain time segment to inform the prediction model for the neighboring time segments.

$$M(l,t) = \sum_{t' \in T} M(l,t') \cdot e^{-\frac{(t-t')^2}{(2 \cdot \sigma^2)}} \quad (3)$$

Equation 3 shows the application of the kernel smoothing to the probabilistic *life-pattern* matrix. We choose gaussian kernel to simulate the temporal decay of the location likelihood based on observations. The choice this kernel is based on three characteristics of the gaussian function:

- symmetric,
- smooth, and
- exponential.

We choose a symmetric function to allow an observation at a certain time segment to equally inform the prediction model for the past time segments as it does for the future time segments. This equality is, as described in the example above, based on the belief that, based on the observation that the mobile user was in Queens at 08:00, it is as logical to assume the user was in Queens at 07:50 as to anticipate the user will continue to be in Queens at 08:10.

We choose a smooth function to simulate the continuous nature of the temporal transition in a mobile user's position. That is, people do not suddenly disappear from one location and reappear at a remote location instantaneously. Therefore, an observation of a mobile user being in Queens at 08:00 can indicate that the mobile user was in Queens during a time period which includes 08:00. Hence, the probabilities in the *life-pattern* matrix should change gradually over time.

We choose an exponential function based on the conventional notion that unless there is an actual model which simulates the temporal dynamics of the data, the confidence of an observation decays exponentially over time [9]. Since the observations available in anonymized CDRs are irregular and can not exactly capture the actual time segment in which the mobile user moves, estimating a prototypical model for simulating the temporal decay of the likelihood in the mobile user's position is not feasible. Therefore, we employ the convention of an exponential decay function.

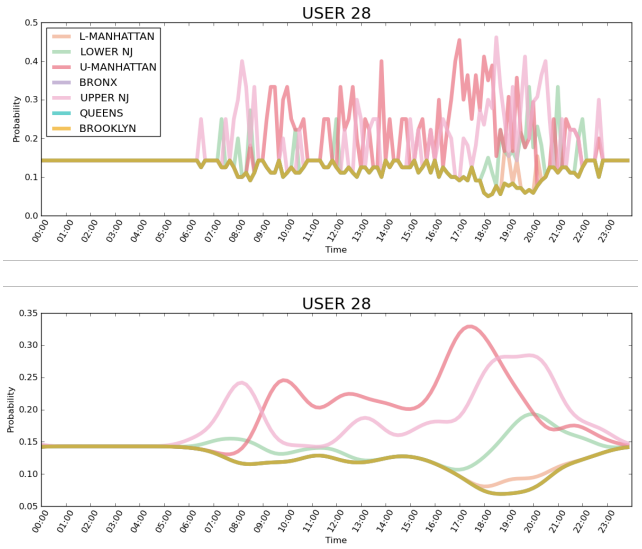

Figure 3: Line charts of a mobile user's probabilistic *life-pattern*. Top is the *life-pattern* without smoothing, and the bottom is the *life-pattern* with smoothing. With the help of the smoothing function, it is clear that the mobile user commutes from Upper New Jersey to Uptown Manhattan.

Figure 3 shows two line charts which represent a mobile user's probabilistic *life-pattern*. The line chart on the top represents an example without smoothing, and the line chart on the bottom represents the same example with smoothing. It is clear from these examples that smoothing allows a more gradual transition from one area to another, while the original *life-pattern* shows little consistency in the mobile user's behavior.

In our analysis, we refer to the temporal sequence of the region which contains the highest likelihood at each time segment as the mobile user's daily trajectory. For example, the daily trajectory of the mobile user shown in figure 3 starts from Upper New Jersey to Uptown Manhattan at 09:00, then back to Upper New Jersey around 18:30.

4.3 Location Predictability

Although the model we propose in Section 4.2 succeeds in extracting probabilistic *life-patterns* of many mobile users, some mobile users have less predictable *life-patterns* than others. This is because not all mobile users have consistent daily routines. Also, not all mobile users provide enough anonymized CDRs for our model to achieve a good approximation of *life-patterns* within the duration of our observation (62 days).

The predictability of a *life-pattern* approximation can be

represented by a series of entropies[3].

Equation 4 represents the calculation of a mobile user's *life-pattern* entropy at time segment t. In our work, we also take the average of each mobile user's temporal entropy values over time to quantify the mobile user's overall predictability.

$$H(t) = -\sum_{l \in L} M(l,t) \cdot logM(l,t) \quad (4)$$

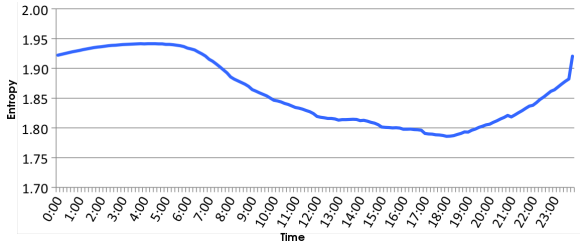

Figure 4: The average entropy of all *life-pattern* matrices over time. During midnight and 06:00 most people were asleep, and there were not many phone calls; hence, the entropy of mobile user *life-pattern* is high. Many people made phone calls around 18:00 as they left work; hence, the entropy is low.

Figure 4 shows a line chart of the average entropy of all *life-pattern* matrices derived from the New York area anonymized CDRs. From this line chart, we can observe that the average entropy starts to drop around 06:00 and decreases throughout the day until the turning point at 18:00. According to domain experts, 18:00 is when many people make phone calls as they leave work. This may explain the predictability peak at 18:00. After 18:00, the average entropy increases until midnight and is stable to the morning. Since most people do not make phone calls during the night, high entropy during these hours is anticipated.

Figure 5 shows a scatter plot of mobile users and a histogram showing the distribution of their overall entropies. The scatter plot suggests that while there is a negative correlation between the mobile users' phone usage and their entropies, the variance in their entropies also grows as more observations are made. Therefore, we can assess that although the predictability of the *life-pattern* of a mobile user increases with communication usage, the user's overall predictability depends more on the regularity of his/her daily routines than the number of observations. Such mobile users whose predictabilities are low despite of their frequent usage are represented by the points above the top orange line in the scatter plot.

The histogram in figure 5 shows that most mobile users in our data have high entropy. This indicates that most users are not sufficiently predictable. There are two main reasons for a mobile user to have such high entropy. The first is due to the lack of observations available in the data. This observation shortage can be simply because these mobile users actually seldom make phone calls, or although their billing addresses are within New York, their actual *life-patterns* take part in other areas. The second is due

[3]The lower the entropy, the more predictable the user's location [25].

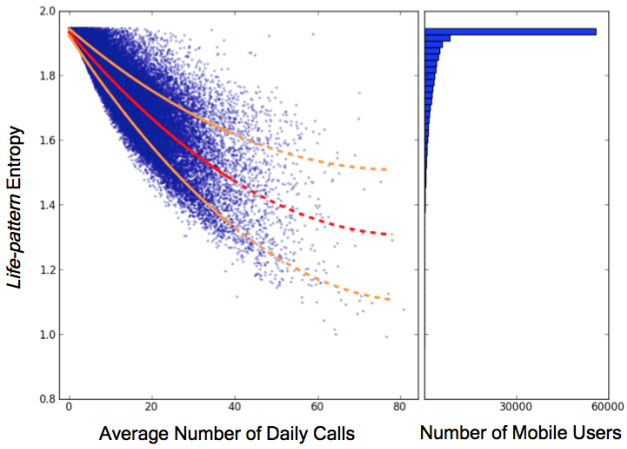

Figure 5: A scatter plot of mobile users and a histogram. Each point in the scatter plot represents a mobile user whose x-coordinate corresponds to the user's frequency of calls, and y-coordinate represents the average entropy over all time segments in the user's *life-pattern*. The red and orange lines indicate the mean entropy over the mobile users with same call frequency and its standard deviation. These lines are modeled using second-order least squares method over the actual values.

to their irregular *life-patterns*, as discussed above, that our model could not extract any consistencies in their mobility habits.

In the following case study, we remove users with high entropy. Although this filtering may seem to contradict to our statement of including mobile users with few phone calls into the analysis, it allows us to focus on mobile users whose lives are deeply rooted into the area and to eliminate the occasional visitors from disrupting the analysis. In other words, this filtering allows us to effectively remove mobile users whose billing addresses are registered to New York however are actually living in other areas. Therefore, the resulting data is fundamentally different from datasets which are filtered based on the phone call frequencies, and it retains many mobile users who dispite of their limited phone call activities have highly structured *life-patterns*.

5. A CASE STUDY

In this section, we demonstrate our analysis on human movements using the information extracted from the New York anonymized CDRs based on the methods we introduced in section 4. The analysis uses several visualizations to help explain the information including movement patterns and the implications of geographical features.

Understanding spatial distribution of population and its temporal dynamics can provide analysts with an overview of large population's movement patterns in urban areas. We have developed a system that allows analysts to explore our models. Our system uses two visual representations to depict collective mobility trends. One visualizations focuses on the spatial distribution of the population, and the other focuses on temporal dynamics of inter-regional traffic.

The spatial distribution is depicted with a node-link diagram drawn on a map. This representation of collective

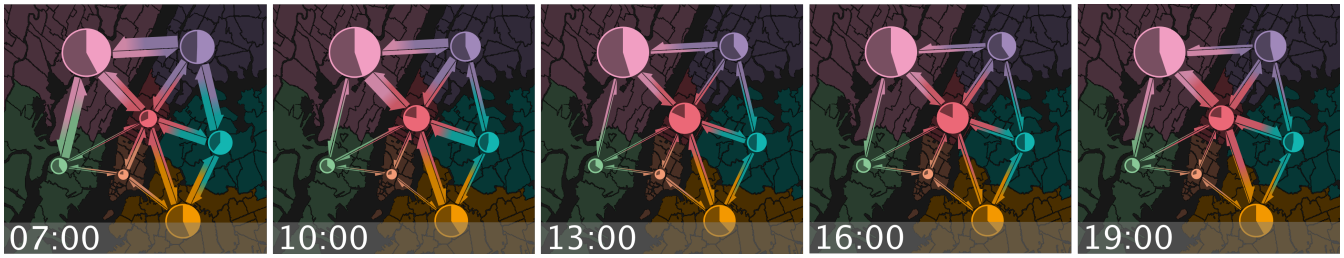

Figure 6: Migration toward Uptown Manhattan significantly grows during the morning causing its node to expand. However, this trend reverses at 19:00, and more people start to migrate out of Uptown Manhattan.

human movement data is applied in many visual analytics systems [4]. The size of a node represents the population of the corresponding region. The width of an edge represents the size of traffic between the endpoint regions that take part during the time segment (10 minutes). This visualization can also be animated through the day by using a time slider interface.

The temporal dynamics of the inter-regional traffic is depicted in two stack graphs. One stack graph shows the temporal fluctuation of the people flowing into the region, and the other shows the people flowing out of the region. This separation of flows allows analysts to observe overall traffic between two regions and also understand region-specific traffic trends. These stack graphs are combined into a single view. In this view, the horizontal line in the middle is the time axis, and inflow stacks are stacked vertically upward, and outflow stacks are stacked vertically downward. In the inflow stacks, the stacks are color-coded to the region of the mobile users' origin, and in the outflow stacks, the stacks are color-coded to the region of the migrating mobile users' destination. The width of each stack corresponds to the migrating population.

Figure 6 shows a set of visualizations representing the spatial distributions of mobile users in New York City. In these visualizations, each node is also presented as a pie chart showing the overall proportion of migrating people and non-migrating people. The dimmed area in each pie chart represents the proportion of mobile users that do not migrate throughout the day.

For example, the node representing Uptown Manhattan contains almost a quarter of dimmed area. This implies that more than three quarters of the population in Uptown Manhattan consist of people who have migrated into Uptown Manhattan or people who will be migrating out of Uptown Manhattan at some point of the day. This information indicates the fluidity of the population in different areas and allows analysts to better understand the characteristics of each region.

In addition to the spatial distribution of population, this visualization also allows analysts to view mobility trends at selected time segments. This information is shown in the edges that connect the nodes. There are two edges for each pair of nodes that represent spatially adjacent regions. Each edge has a direction, which indicates the direction of migration, and its width represents the amount of traffic. In figure 6, we can see that the traffic between Uptown Manhattan and other regions concentrates towards Uptown Manhattan during the morning, and this trend reverses in the night. This mobility trend matches rush hour in New York City and confirms the mobile users' *life-patterns* who commute to Uptown Manhattan.

One anomaly we found here is that despite of the fact that Downtown Manhattan contains the Financial District, one of the busiest areas in Manhattan, the population of Downtown Manhattan remains low throughout the day. Although this may require further investigation into the actual data, we consider this may be caused by the initial preparation of removing business owned cellular phones from the anonymized CDRs.

Figure 7 shows three views of mobility stack graphs. Each stack graph is augmented with an additional line chart, which indicates the overall fluctuation of the population in the focus region. This line chart allows analysts to easily assess the overall flux of the migrations.

Comparing the line charts of the overall mobile user population in Uptown Manhattan and Queens, we can see that while the number of distinct mobile users in Uptown Manhattan increases during the day, this number decreases in Queens during the day. This implies that Uptown Manhattan is more likely to be a business area, where people gather to work during the day, while Queens is more likely to be a residential area, where people leave in the morning and return at night. Although the number of mobile users in Upper New Jersey remains relatively stable throughout the day, it is evident from the stacked graph that there is a significant number of mobile users entering and exiting this region. Therefore, it appears that Upper New Jersey has both business and residential areas.

The stack graphs in figure 7 also suggest that mobile user migration is least active between 12:00 and 16:00. Since this study was limited to mobile users with highly structured *life-patterns*, who are likely to have regular jobs, apparently most of these mobile users stay in their offices and do not migrate during this time period.

In addition to this common feature, the unique geometry of the Uptown Manhattan stack graph also conveys interesting information. For example, the steady growth of the incoming population peaking around 09:00 aligns with rush hour in New York City. Another interesting trend is that Uptown Manhattan seems to have two peaks for outgoing traffic. One is at around 18:00 and the other is around 04:00. While the 18:00 peak is expected, as the closing time of most day jobs are at around 17:00, the peak at 04:00 does not have an obvious reason. However, under the New York state law, bars are required to stop serving alcohol by 04:00. Therefore, although this requires further investigation, it may be that the outgoing peak around 04:00 in Upper Manhattan is based on the people who leave the city when the bars close.

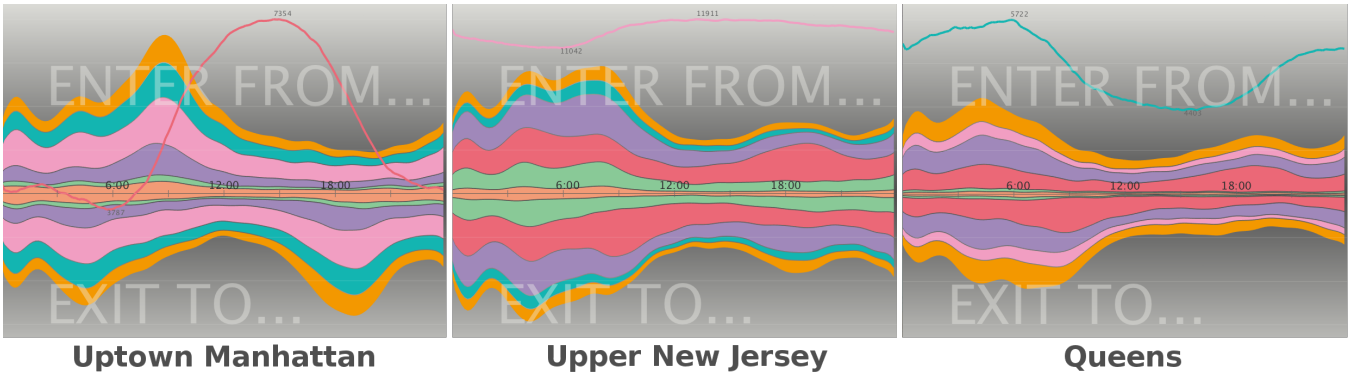

Figure 7: Stack views of Uptown Manhattan, Upper New Jersey, and Queens. Each stack view is augmented with a line chart showing the fluctuation of overall population in the region. Each stack view shows the region's characteristic based on the mobility patterns of mobile users.

6. CONCLUSIONS

In this paper, we introduce a new method for inferring human movement patterns from anonymized CDRs. Our method focuses on approximating the *life-patterns* of mobile users within a geographic context. The model we propose in this paper copes with sparsity in mobile users' phone usage by combining observations in anonymized CDRs from different days, and therefore enables analysis on a wider range of subjects than previous studies that focus only on heavy users.

Our study also includes a new algorithm for partitioning maps into spatially contiguous community regions and contains a metric for quantifying the predictability of a mobile user's *life-pattern*. In our experiment, the map partitioning algorithm was applied to movements between zip codes. The results from this experiment in section 4.1 suggest that not only do geographical features constrain human movement, but demographical and other features also potentially have an effect on human mobility that can be seen in the data. The entropy-based metric showed that although there is a significant correlation between the frequency of communication events and the predictability of *life-patterns*, beyond a certain level predictability is more dependent on regularity in a subject's daily routine.

In order to demonstrate our method, we conducted a case study of analyzing human mobility in New York City and its surrounding areas. In this case study, several widely known regional characteristics as well as new characteristics were conveyed through a series of visualizations.

Although the set of procedures we introduce in this paper have yet to undergo further experimentation to examine its correctness and applicability under various situations, our case study on New York City anonymized CDRs show significant potential in revealing contextual information of human mobility regarding geographic, demographic, and other characteristics of the city. We believe our method provides a foundation for applying anonymized CDRs to analyze human mobility with respect to its geographic context and facilitate practical analyses of urban mobility while maintaining mobile user anonymity. Our future work will focus on further investigating the credibility of *life-patterns* extracted by our method in concern with the observation period of anonymized CDRs. We also plan to conduct a comparative study on human mobility in different cities to better understand how geography and other contextual underlying information affect human mobility and the characteristics of urban cities.

7. REFERENCES

[1] T. Anderson, V. Abeywardana, J. Wolf, and M. Lee. National travel survey gps feasibility study final report. Technical report, Department of Transport, Dec 2009.

[2] G. Andrienko, N. Andrienko, P. Bak, S. Bremm, D. Keim, T. von Landesberger, C. Pölitz, and T. Schreck. A framework for using self-organizing maps to analyze spatio-temporal patterns, exemplified by analysis of mobile phone usage. *Journal of Location Based Services*, pages 200–221, 2010.

[3] G. Andrienko, N. Andrienko, F. Giannotti, A. Monreale, and D. Pedreschi. Movement data anonymity through generalization. In *Proc. 2nd SIGSPATIAL ACM GIS 2009 International Workshop on Security and Privacy in GIS and LBS*, pages 27–31, 2009.

[4] N. Andrienko and G. Andrienko. Spatial generalization and aggregation of massive movement data. *Trans. IEEE Visualization and Computer Graphics*, 17:205–219, 2011.

[5] R. Becker, R. Caceres, K. Hanson, J. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky. A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing*, 10(4):18 –26, Apr 2011.

[6] S. Boettcher and A. G. Percus. Optimization with extremal dynamics. *Complex.*, 8:57–62, Nov 2002.

[7] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner. On modularity clustering. *Trans. IEEE Knowledge and Data Engineering*, 20(2):172 –188, Feb 2008.

[8] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439:462–465, 2006.

[9] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer-Verlag, New York, NY, 2009 edition, 2009.

[10] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, Feb 2010.

[11] A. Frihida, D. J. Marceau, and M. Thériault. Spatio-temporal object-oriented data model for disaggregate travel behavior. *Trans. GIS*, pages 277–294, 2002.

[12] T. Fujisaka, R. Lee, and K. Sumiya. Exploring urban characteristics using movement history of mass mobile microbloggers. In *Proc. 11th ACM Workshop on Mobile Computing Systems and Applications (HotMobile)*, pages 13–18, 2010.

[13] F. Girardin, A. Gerber, C. Ratti, A. Vaccari, and A. Biderman. Towards estimating the presence of visitors from the aggregate mobile phone network activity they generate. In *Proc. ACM International Conference on Computers in Urban Planning and Urban Management*, pages 52–61, 2009.

[14] M. C. González, C. A. Hidalgo, and A.-L. BarabÃąsi. Understanding individual human mobility patterns. *Nature*, 453:779–782, Jun 2008.

[15] D. Guo. Visual analytics of spatial interaction patterns for pandemic decision support. *International Journal of Geographical Information Science*, 21:859–877, January 2007.

[16] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, J. Rowland, and A. Varshavsky. A tale of two cities. In *Proc. 11th ACM Workshop on Mobile Computing Systems and Applications (HotMobile)*, pages 19–24, 2010.

[17] B. W. Kernighan and S. Lin. An Efficient Heuristic Procedure for Partitioning Graphs. *The Bell system technical journal*, 49(1):291–307, 1970.

[18] C. Y. Ma, D. K. Yau, N. K. Yip, and N. S. Rao. Privacy vulnerability of published anonymous mobility traces. In *Proc. 16th annual international conference on Mobile computing and networking*, pages 185–196, New York, NY, USA, 2010. ACM.

[19] M. Martino, F. Calabrese, G. Di Lorenzo, C. Andris, L. Liang, and C. Ratti. Ocean of information: fusing aggregate & individual dynamics for metropolitan analysis. In *Proc. 15th ACM International Conference on Intelligent user interfaces (IUI)*, pages 357–360, 2010.

[20] A. M. Matthew Block, Shan Carter. Mapping america: Every city, every block, Nov 2011.

[21] R. P. Minch. Privacy issues in location-aware mobile devices. In *Proc. 37th Annual Hawaii International Conference on System Sciences*, Washington, DC, USA, 2004. IEEE Computer Society.

[22] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *PHYS.REV.E*, 69:066133, 2004.

[23] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, Feb 2004.

[24] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti. Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing*, 6:30–38, 2007.

[25] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423,623–656, 1948.

[26] C. Song, T. Koren, P. Wang, and A.-L. Barabasi. Modelling the scaling properties of human mobility. *Nature Physics*, 6:818–823, Sept. 2010.

[27] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of Predictability in Human Mobility. *Science*, 327:1018–1021, Feb. 2010.

[28] United States Census Bureau. American community survey, Nov 2011.

[29] M. R. Vieira, E. Frías-Martínez, P. Bakalov, V. Frías-Martínez, and V. J. Tsotras. Querying spatio-temporal patterns in mobile phone-call databases. In *Proc. 11th IEEE International Conference on Mobile Data Management (MDM)*, pages 239 –248, May 2010.

[30] M. P. Wand and M. C. Jones. *Kernel Smoothing (Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, 1 edition, Dec. 1994.

[31] H. Zang and J. Bolot. Anonymization of location data does not work: a large-scale measurement study. In *Proc. 17th annual international conference on Mobile computing and networking*, pages 145–156, New York, NY, USA, 2011. ACM.