Identifying Important Places in People's Lives from Cellular Network Data

Sibren Isaacman[1], Richard Becker[2], Ramón Cáceres[2], Stephen Kobourov[3], Margaret Martonosi[1], James Rowland[2], and Alexander Varshavsky[2]

[1] Dept. of Electrical Engineering, Princeton University, Princeton, NJ, USA
{isaacman,mrm}@princeton.edu
[2] AT&T Labs – Research, Florham Park, NJ, USA
{rab,ramon,jrr,varshavsky}@research.att.com
[3] Dept. of Computer Science, University of Arizona, Tucson, AZ, USA
kobourov@cs.arizona.edu

Abstract. People spend most of their time at a few key locations, such as home and work. Being able to identify how the movements of people cluster around these "important places" is crucial for a range of technology and policy decisions in areas such as telecommunications and transportation infrastructure deployment. In this paper, we propose new techniques based on clustering and regression for analyzing anonymized cellular network data to identify generally important locations, and to discern semantically meaningful locations such as home and work. Starting with temporally sparse and spatially coarse location information, we propose a new algorithm to identify important locations. We test this algorithm on arbitrary cellphone users, including those with low call rates, and find that we are within 3 miles of ground truth for 88% of volunteer users. Further, after locating home and work, we achieve commute distance estimates that are within 1 mile of equivalent estimates derived from government census data. Finally, we perform carbon footprint analyses on hundreds of thousands of anonymous users as an example of how our data and algorithms can form an accurate and efficient underpinning for policy and infrastructure studies.

1 Introduction

While people travel further and faster than ever before, it is still the case that they spend much of their time at a few important places. Identifying these key locations is thus central to understanding human mobility and social patterns. Such understanding can, in turn, inform solutions to large-scale societal problems in fields as varied as telecommunications, ecology, epidemiology, and urban planning. As an example, knowing how large populations of people move about would help determine their carbon footprint and in turn help guide policies intended to reduce that footprint.

Wireless cellular networks hold great potential for providing the necessary information to identify important places in people's lives. The growing ubiquity of cellular phones means that a large percentage of people keep a phone with them most of the time. In addition, the networks need to know roughly where each phone is in order to provide the phones with voice and data services.

In this work, we explore the use of anonymized Call Detail Records (CDRs) from a cellular network to estimate the locations of important places in the lives of large

Proc. of 9th International Conference on Pervasive Computing (Pervasive), June 2011

populations of people. CDRs are routinely collected by cellular network providers to help operate their networks, for example to identify congested cells in need of additional bandwidth. Each CDR contains information such as the time a voice call was placed or a text message was received, as well as the identity of the cell tower with which the phone was associated at that time. This information can serve as sporadic samples of the approximate locations of the phone's owner.

CDRs are an attractive source of location information for two main reasons. One, they are collected for all active cellular phones, which number in the hundreds of millions in the US and in the billions worldwide. Two, they are already being collected to help operate the networks, so that additional uses of CDR data incur little marginal cost. Contrast this low cost, for example, with the expense of carrying out surveys to ask people where they spend their time. This high expense generally limits other data collection methods to orders of magnitude fewer participants.

However, CDRs have two significant limitations as a source of location information. One, they are sparse in time because they are generated only when a phone engages in a voice call or text message exchange. Two, they are coarse in space because they record location only at the granularity of a cell tower. It is an interesting research question whether CDRs can be used to identify important places in people's lives.

In this paper, we show that applying clustering and regression techniques to CDR data can indeed identify important places in people's lives. First, we present an algorithm for identifying important places. Then, we describe two other algorithms for selecting home and work locations from among those important places. We validate all three algorithms by comparing their results to ground truth provided by a group of volunteers. We then apply these algorithms to much larger anonymous populations in the Los Angeles (LA) and New York City (NY) areas. Our LA and NY dataset spans two months of activity for hundreds of thousands of phones, yielding hundreds of millions of location samples.

Finally, we present two example applications of these techniques. We start by using the home and work locations identified by our algorithms to calculate the distribution of commute distances per postal code in our Los Angeles and New York dataset. We then estimate the carbon footprints of those commutes, also aggregated by postal code.

Overall, the contributions of our work are as follows.

- We propose and evaluate a model based on logistic regression of volunteers' locations for *Important Places* analysis. In our first algorithm, we demonstrate an accurate and efficient method for identifying *Important Places* from CDRs. Our algorithm is the first to operate on the majority of cellular phone users, rather than relying either on more continuous and fine-grained tracking (e.g. GPS) or focusing on high-call-rate users whose mobility is easier to analyze.
- We propose and evaluate two other algorithms for applying semantic meaning to important locations, namely *Home* and *Work*, using other models also derived via logistic regression. Our algorithms identify these key sites with median errors under one mile.
- We test our approaches on a dataset that is more universal than prior work in several ways. First, it is simply larger than prior work in terms of CDRs and number of users. Second, it covers multiple distinct geographic areas. Third, it considers users

with a wide variety of call/text rates, from as low as a few calls/texts per week up to dozens of calls/texts per day.

– Finally, we provide examples of how technology providers and policy makers might use our data and algorithms in their work. In particular, we calculate home-to-work commute distances and combine them with publicly available data to estimate the carbon footprints of those commutes in two major metropolitan areas. Our average commute distances for the LA and NY areas are within 1 mile of the equivalent averages computed from US Census data.

In summary, our work extends prior research in location identification and cellphone mobility to create effective algorithms and solid foundations for technological and societal problem-solving. The rest of this paper is organized as follows. Section 2 describes the data we obtained from volunteers as well as the much larger set of anonymous CDRs, including the measures we have taken to preserve individual privacy. Section 3 presents our algorithm for identifying important locations, Section 4 our algorithms for selecting home and work locations, and Section 5 our estimates of commuting distances and carbon footprints. Section 6 discusses related work, and Section 7 offers conclusions.

2 Data Collection Methodology and Characteristics

2.1 Anonymized Call Detail Records

We collected anonymized Call Detail Records (CDRs) from a random set of cellular phones whose billing addresses lie within specific geographic regions.

Defining Geographic Regions of Interest: We first developed a target set of 891 postal (ZIP) codes located in the Los Angeles and New York metropolitan areas. In the LA area, the ZIP codes cover the counties of Los Angeles, Orange, and Ventura. In the NY area, these ZIP codes cover the five New York City boroughs (Manhattan, Brooklyn, Bronx, Queens, and Staten Island) and ten New Jersey counties that are close to New York City (Essex, Union, Morris, Hudson, Bergen, Somerset, Passaic, Middlesex, Sussex, and Warren). Figure 7 shows maps of the regions studied as part of carbon footprint results presented in Section 5. Our selected ZIP codes cover similarly sized areas in LA and NY.

Anonymized CDR Contents: We then obtained anonymized CDRs for a random sample of phones in each ZIP code. The CDRs contain information about two types of events involving these phones: voice calls and text messages. In place of the phone number, each CDR contains an anonymous identifier consisting of the 5-digit billing ZIP code and a unique integer. Each CDR also contains the starting time of the voice or text event, the duration of the event, the locations of the starting and ending cell towers associated with the event, and an indicator of whether the phone was registered to an individual or a business. It is important to note that we collect CDRs for these phones wherever in the US they travel, not only when they contact cell towers within their billing ZIP codes.

Excluded Categories of Phones: Our goal is to understand aggregate mobility patterns of people in particular regions of the country, and to compare them analytically

where possible. As such, our study omits from consideration two sets of phones from the original CDRs.

First, we omitted phones registered to businesses, retaining only phones registered to individuals. This step avoids, for example, the situation where a cellular service reseller based in a ZIP code of interest would cause us to study large numbers of phones that are not representative of that ZIP code.

Second, we removed from our sample those phones that appeared in their base ZIP code fewer than half the days they had voice or text activity. We assumed that the owners of such phones now live in other parts of the country but have retained their old billing addresses (e.g., they are college students). Therefore, their daily travel patterns may not be representative of the geographical areas we are interested in.

After these two filtering steps, our CDRs are a useful representation of mobility and telephone usage in the regions of interest. While there will always be some people using personal phones for business (and vice versa), we have compared our filtered CDRs against US Census data for the regions of interest [22] and found a strong correlation between the expected and actual number of users in each ZIP code.

Dataset Characteristics: Our data collection methodology resulted in location data for hundreds of thousands of phones split roughly evenly between LA and NY, with the number of phones in each ZIP code proportional to the population in that ZIP code. We collected data for 78 consecutive days from November 15, 2009, to January 31, 2010. Table 1 offers some general characteristics of this dataset. As shown, it contains hundreds of millions of location samples, with on the order of 10 location samples per phone per day.

Metric	LA	NY
Total Unique Phones	97K	71K
Total Unique CDRs	247M	161M
Median Calls Per Day	8	9
Median Texts per Day	4	3

Table 1. General characteristics of our Call Detail Record dataset.

Privacy Measures: Given the sensitivity of the data, we took several steps to ensure the privacy of individuals represented in our CDR dataset.

First, only anonymous records were used in this study. In particular, personally identifying characteristics were removed from our CDRs. CDRs for the same phone are linked using an anonymous unique identifier, rather than a telephone number. No demographic data is linked to any user or CDR.

Second, all our results are presented as aggregates. That is, no individual anonymous identifier was singled out for the study. By observing and reporting only on the aggregates, we protect the privacy of individuals.

Finally, each CDR only included location information for the cellular towers with which a phone was associated at the beginning and end of a voice call or at the time of a text message. The phones were effectively invisible to us aside from these events. In addition, we could estimate the phone locations only to the granularity of the cell tower

coverage radius. Although the effective radius depends upon tower height, radio power, antenna angle, and terrain, these radii average about a mile, giving an uncertainty of about 3 square miles for any event [21].

2.2 Ground Truth Data from Volunteers

In order to validate our work, we recruited a group of 37 volunteers who provided us the true locations of important places in their lives, as well as permission to inspect their CDRs for the purposes of this study. The group is composed of graduate students and professionals, all of which are personal or professional acquaintances of the authors. Of the 37 volunteers, 29 are male and 8 are female. Geographically, 31 recruits live in the states of New York or New Jersey and the remaining 6 live in Ohio, Georgia, or Arizona. The majority of the volunteers work at high-tech jobs.

Each volunteer filled out a survey on a website. The survey form asked them to list up to 10 important places in their lives, defined as places where they had spent a significant amount of time and/or visited frequently in the previous 60 days. It specifically requested that they include home and work in the list, and expressed the hope that they would list additional places such as a gym or the destination of an overnight trip.

The volunteers also provided us with the latitude and longitude of each place they listed. The survey website included a tool to help them find this information. Volunteers could either enter a street address, or drop a pin on a map after panning and zooming the map to the appropriate location. The tool would convert this input into a latitude-longitude pair that the volunteer could cut and paste into the survey form.

In the work described in the rest of this paper, we used the ground truth data from 18 of our volunteers as a training set for our algorithms, and data from the remaining 19 volunteers as a testing set. The 18 training volunteers were chosen arbitrarily and without regard for their mobility or calling patterns. For both our training and testing volunteers, we collected CDRs for the same 60 days covered by their survey responses.

3 Identifying Important Places

Intuitively, we know that human mobility involves moving to and from a set of places, some of which are recurringly important to us and some of which are visited less often or only fleetingly. Being able to discern significant places in people's lives is an important aspect of characterizing human mobility. Identifying important places can be used to support location-based services, improve understanding of general human movement patterns, and support the creation of realistic and practical models of human mobility. We define an *important place* as a geographic location where a person spends a significant amount of time and/or which she visits frequently. Examples of important places include: home, work, gym, grocery store, and a house of worship.

In this section, we show how mobile network events can be used to identify important places in people's lives. We identify important places based on the mobile network *events* that correspond to CDR entries. Thus, making or receiving a phone call or sending or receiving a text message generates an *event*. For each event (CDR), we know its time of occurrence and the location of the first and last cell towers associated with it. We refer to the list of events that were generated by a user's phone as the user's *trace*. If a cell tower appears in the user's trace on a given day, we say that the cell tower was *contacted* on that day.

Our algorithm for identifying important places has two stages. In the first stage, we spatially cluster the cell towers that appear in a user's trace. In the second stage, we identify which of the clusters are important using a model derived from a logistic regression of volunteers' CDRs. In the rest of this section, we describe these two stages in detail, present our validation results based on the important locations of our 19 testing volunteers, and compare the results characteristics for our NY and LA populations.

While in this paper we use cellular phone activity collected in the form of CDRs, our algorithms for identifying important locations, and for assigning semantic meaning to these locations such as "home" or "work", are quite general. They could also be applied to location traces collected via GPS, WiFi, or other techniques.

3.1 Clustering Cell Towers

Clustering cell towers that appear in a user's trace has two steps. In the first step, we sort the cell towers in descending order based on the total number of days they were contacted. Thus, the cell tower that was contacted on the most days will be ranked first. Sorting the cell towers serves as a modest optimization but is not required. An on-line algorithm could easily be developed by removing the sorting phase, resulting in an average change in error of less than 0.1 miles. However, sorting by the number of days the cell tower was contacted ("call-days") rather than by the total number of events associated with the cell tower is both important and novel. In particular, sorting by call-days rather than total calls helps to decrease the influence of cell towers that were contacted only on a few days, but that had a burst of events on those days. A flurry of calls from one location on a single day is not as indicative of location importance as a similar number of calls spread over many days at a location that recurs. Consider, for example, work travel to a distant location. Though the trip may be short in duration, one might make many calls back home to family and friends. These calls would then unduly increase the perceived importance of the location. This distinction helps us to maintain good location accuracy for users across a wide range of calls-per-day.

In the second step, the sorted list of cell towers is clustered according to Hartigan's leader algorithm [10]. We chose the leader algorithm because it doesn't require pre-specifying the desired number of clusters and because it works in a single pass, which is important for practical use on very large datasets such as ours (4GB, compressed).

The leader algorithm starts with the first cell tower in the sorted list and makes this tower the centroid of the first cluster. Then, for each subsequent cell tower, it checks to see whether the tower falls within a threshold radius of the centroid of any existing cluster. If it does not, the tower becomes the centroid of a new cluster. If it does fall within the threshold radius of an existing cluster, the algorithm adds the tower to the cluster and moves the centroid of the cluster to be the weighted average of the locations of all the cell towers in the cluster. The cell tower locations, in our case, are weighted by the number of call-days. The algorithm completes once every cell tower has been assigned to a cluster.

Choosing a particular threshold radius around cell towers helps equalize for the fact that in urban areas towers might be as dense as 200 meters apart, while in suburban areas, spacings of 1-3 miles are more common. We experimented with a range of radii and found that 1 mile works well in practice.

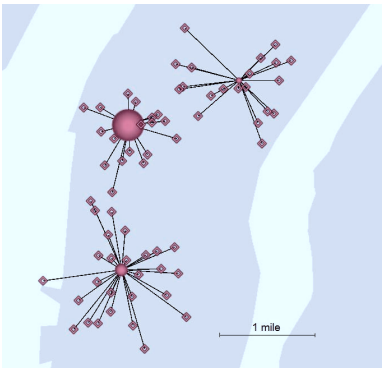


Fig. 1. Our clustering algorithm applied to a volunteer. Cell towers are clustered into groups according to Hartigan's leader algorithm. Cell towers are added to a cluster if they are within a mile of the cluster's centroid. Clusters are depicted as circles and cell towers as diamonds. A line connects each cell tower to the centroid of its cluster. Circle size is proportional to the number of days on which any cell tower in the cluster was contacted.

Figure 1 illustrates the result of running the clustering algorithm on a volunteer's trace with a threshold radius of 1 mile. We can see that although the volunteer connected to the network through many cell towers, there are only three clusters. Note again that the size of a cluster is proportional to the number of days on which any cell tower in the cluster was contacted, and not necessarily proportional to the number of cell towers that belong to the cluster. For instance, the middle cluster in Figure 1 is the largest even though it contains fewer cell towers than the southernmost cluster. Although for this volunteer there are many cell towers belonging to each of the clusters, it is common for people to have clusters comprising only one or two cell towers.

3.2 Determining Importance

Clustering cell towers typically results in dozens to hundreds of clusters, most of which may have little importance to the user. In this section, we describe how our algorithm determines which clusters are important.

We developed an algorithm for identifying important clusters by studying the behavior of our 18 training volunteers and then testing the algorithm on a set of 19 testing volunteers. Studying the data of our training volunteers revealed the following five observable factors that are considered in determining whether a cluster is important:

- *Days*: The number of days on which any cell tower in the cluster was contacted. If two or more cell towers were contacted on the same day, the day is counted only once. This factor gives a sense of the regularity of activity in the cluster.
- *Tower Days*: The sum of the number of days cell towers in the cluster were contacted. Thus, each cell tower in the cluster adds its *Days* value to the sum. This factor gives a sense of both the number of cell towers in the cluster as well as the number of days on which cell towers in the cluster were contacted.
- *Duration*: The number of days that elapse between the first contact with any cell tower in the cluster and the last contact with any cell tower in the cluster. (For example, a cluster with one cell tower that was contacted only on the first day and

on the seventh day of the user's trace has a duration value of 6.) *Duration* gives a sense of how long a user is in the area of the cluster, even if network events were not generated from this cluster on a daily basis.

- *Work Hour Events* : The number of times any cell tower in the cluster was contacted on weekdays between 1pm and 5pm. We experimented with various ranges of hours and found that 1pm to 5pm works well in practice because it is a core set of hours for both early and late workers.
- *Home Hour Events* : The number of times any cell tower in the cluster was contacted on weekends or weekdays between 7pm and 7am.

The algorithm identifies a cluster as important if the cluster satisfies each of the following three conditions. First, cell towers in the cluster should have been contacted on more than 5% of the total days in the study. In our case, this translates to the *Days* factor being higher than 2. This condition filters out transitional clusters that are rarely contacted. Second, the cluster should have a *Duration* of more than 14. This helps to remove vacations and other locations that may generate a large number of events in a short period of time but that are not consistently used throughout the trace. Third, the cluster should have a higher than 20% chance of being important according to the regression analysis discussed below. While we derived all the thresholds experimentally based on the data from the 18 training volunteers, our tests on other volunteers and on the larger dataset point to their general applicability.

To determine the likelihood of a cluster being important, we use logistic regression. We considered the five observable factors described above as well as several derived variables. Specifically, we added the rank and the percentage of each of the observable factors. Rank is calculated by ordering the clusters based on the observable factor and then assigning each cluster a sequential number. For example, the cluster with the largest *Duration* gets a ranking of 1 and the cluster with the second largest *Duration* gets a ranking of 2. Percentage is calculated by dividing the value of a given observable factor of a cluster by the sum of these values in all clusters. For instance, if the *Days* value of the current cluster is 5 and there are two more clusters with *Days* values of 20 and 25, the percentage of the *Days* factor of the current cluster is 5 divided by 50, or 0.1. In total, we ended up with 15 observable and derived factors.

$$Prob(x_1, ..., x_n) = \frac{1}{(1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n})} \tag{1}$$

Equation 1 shows the general form of the logistic regression formula that we use to estimate the likelihood of the importance of a cluster. In this formula, $Prob(x_1, ..., x_n)$ is the probability that a cluster with factors x_i is the closest cluster to an important location and β_js are coefficients that are discovered during the regression.

To discover the coefficients, we used the important locations of our 18 training volunteers. First, we marked clusters of each of the volunteers as either being important or not. A cluster is marked as being important if its centroid is the closest to any of the important locations specified by the volunteer. The importance of a cluster is the dependent binary variable in our regression analysis and the 15 observable and derived factors are the independent variables. Once the statistically insignificant factors were eliminated, only three factors were left: the percentage of *Tower Days*, the *Duration*,

and the ranking of a cluster based on *Days*. The percentage of *Tower Days* and the ranking based on *Days* prefers clusters with many cell towers contacted on many days. The *Duration* indicates that for a cluster to be considered important its cell towers must be contacted during a large fraction of the trace. Including the *Duration* feature reduces the importance of transitionary calls made during travel by giving a higher weight to the locations where a user made phone calls many days apart. We conjecture that Duration was selected as one of the main features because people tend to return to places that are important to them often and tend to visit transitionary places infrequently. Once the training is complete, we estimate the importance of a new cluster by feeding these three statistically significant factors of the cluster into the regression formula.

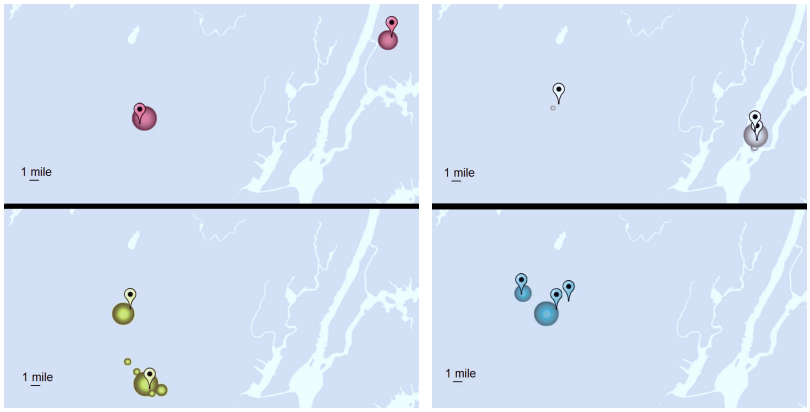


Fig. 2. True important locations vs. discovered important clusters for four volunteers. Paddles represent the important locations provided by the volunteers. Circles represent the important clusters discovered by our algorithm, with their radii signifying days of use. The four examples are drawn to the same scale.

Figure 2 plots the true important locations and the discovered important clusters of four volunteers. The figure confirms that the discovered important clusters match well with the volunteer-provided important locations. However, one important location in the bottom right figure was not matched by any discovered important cluster. This is because the volunteer made almost no calls from that location. It is worth noting that the algorithm performed well despite the significant difference in patterns of of important locations for different volunteers (e.g., different number of important locations, different spatial distributions, different rate of calls, etc.)

3.3 Validation of Important Places Algorithm

We further validate our algorithm for determining important places by comparing to other approaches. Recall that each important cluster contains one or more cell towers. The location of an important place is then the weighted centroid of the geographic locations of these cell towers. To measure how well our *Important Places* algorithm works we calculate the error between each important location provided to us by our 19 testing volunteers and the nearest important place identified by our algorithm. We also compare our results to two additional algorithms: *Nearest Cluster* and *Nearest*

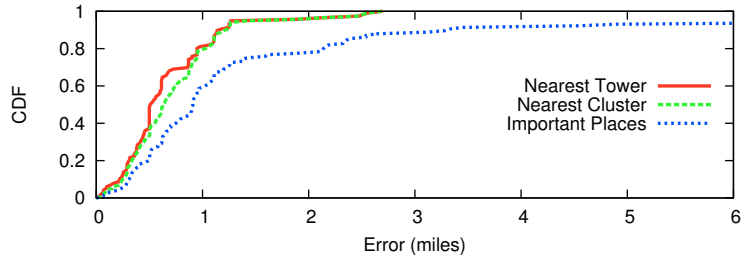
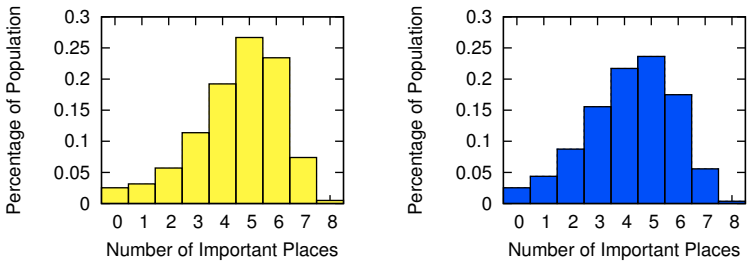


Fig. 3. CDF of errors between true important locations and those found using three techniques. Important Places refers to the clusters identified as important by our algorithm. Nearest Cluster shows the best possible outcome with clustering performed. Nearest Tower demonstrates the best that can be done without clustering.

Tower. The *Nearest Cluster* algorithm considers all clusters identified in Section 3.1 and identifies an important place as the weighted centroid of a cluster that is the nearest to an actual important location. This algorithm acts as an upper bound on our *Important Places* algorithm since it operates on all discovered clusters, without restricting the pool of clusters to choose from. The *Nearest Tower* algorithm determines an important place at the location of the cell tower that is the closest to the actual important location. This algorithm shows the limit of the accuracy we can achieve if we limit our clusters to just a single cell tower.

Figure 3 plots a CDF of the error between the actual important locations and the identified important places for the *Important Places*, *Nearest Cluster* and *Nearest Tower* algorithms. The figure shows that the *Important Places* algorithm performs very well, even though it reduces the total number of identified clusters by up to 90% for some volunteers. The median error of the *Important Places* algorithm is 0.9 miles, which is close to the performance of the upper bound algorithms. *Nearest Cluster* and *Nearest Tower* achieve 0.62 and 0.5 miles median error, respectively. The reason for the slightly higher error for the *Important Places* algorithm is that some people do not use their cell phone frequently at their important locations. Viewing the results more broadly, we find our approach maintains within-3-miles accuracy for 88% of the users, which suffices for the types of policy and planning applications we envision.

3.4 Important Places in Los Angeles and New York



(a) Los Angeles (b) New York

Fig. 4. Histogram of the number of important places of people in Los Angeles and New York.

In this section, we use our full CDR data set to compare the number of important places of Angelenos and New Yorkers, as identified by our *Important Places* algorithm. Without user-provided surveys, we do not know the actual number of important places of people in the LA and NY dataset, but our results allow us to roughly compare the behavior of people in the two areas. Figure 4 plots the histograms of the number of important places for the two populations. We draw three conclusions from our results. First, our algorithm succeeds in finding important locations for 97.5% of the user population, which is a very high percentage. For the remaining 2.5%, our algorithm cannot find any important places, mostly due to people not using their cell phones frequently enough for our algorithm to work. Second, about a quarter of people in both areas have exactly 5 important places and more than three quarters of people have between 3 to 6 important places. This is similar to results found in other work [2, 9]. Third, New Yorkers have a higher percentage of people with 1 to 4 important places, whereas Angelenos have a higher percentage of people with 5 to 8 places. Overall, the implications are that although typical cellphone users have dozens or hundreds of towers contacted in a multi-month event trace, their main locus of mobility is concentrated on a much smaller set of places.

4 Identifying Home and Work

In this section, we move beyond identifying generally important places in people's lives, to inferring where people live and work. The knowledge of where people live and work can be used for a detailed analysis of mobility prediction models, workday patterns, commuter carbon footprints, and a variety of context-aware applications, such as location-based reminders [20].

We developed algorithms that compute estimates of where a cellphone user lives and works, given a list of important clusters identified in Section 3.2. We call these estimated locations *Home* and *Work*, respectively. Of course, not everyone will have distinct home and work locations: some people work at home, others have no fixed work site, and still others may not use their cell phones at home and/or work. Nevertheless, our validation work confirms that our algorithms produce good approximations of the true home and work locations of volunteers.

4.1 Home and Work Algorithms

Our Home and Work algorithms select, among all important clusters identified by the Important Places algorithm described in Section 3.2, the clusters that correspond to where a person lives and works, respectively. The algorithms are independent and may end up selecting the same cluster as both Home and Work.

To select Home or Work, the relevant algorithm (i.e., either the Home or Work algorithm) calculates a score for each important cluster using coefficients obtained from a logistic regression. The algorithm then assigns the cluster with the highest score to be Home or Work. To calculate a score for a cluster, we use the logistic regression formula shown in Equation 1. In this case, $Prob(x_1, ..., x_n)$ is the score calculated for each cluster, x_i is the value of the ith factor and β_js are regression coefficients fitted during training. To train our regression formulae, we repeated the procedure described in Section 3.2 using the reported home and work locations of the 18 training volun-

Percentile	25^{th}	50^{th}	75^{th}	95^{th}
Home Error	0.53	0.90	1.28	3.86
Work Error	0.62	0.83	2.30	21.23

Table 2. Errors in miles from true home and work locations to those found using our Home and Work algorithms.

teers. Home and Work, then, are chosen as the clusters with the highest probability as computed by the coefficients given by the logistic regression.

Recall that in this study we define "home" hours to be weekends and weekdays between 7PM and 7AM, whereas "work" hours are weekdays between 1pm and 5pm. For the Home algorithm, the single most dominating factor was the *Home Hour Events*. That is, the cluster with the largest number of events during the "home" hours is selected as Home. For the Work algorithm, there are two dominating factors. The first factor is the rank of the *Work Hour Events*. In other words, after ranking all clusters based on the number of events occurring during "work" hours, a cluster with a higher ranking is assigned a higher score than a cluster with a lower ranking. The second factor is the percentage of the *Home Hour Events*. Recall that this percentage is calculated as the number of events occurring during "home" hours in the cluster, divided by the total number of events occurring during "home" hours in all clusters. A cluster is assigned a higher score by the Work algorithm if the percentage of the *Home Hour Events* in the cluster is low.

4.2 Validation of Home and Work Algorithms

Table 2 shows the 25^{th}, 50^{th}, 75^{th} and 95^{th} percentile errors between the Home and Work locations as estimated by our Home and Work algorithms and the actual home and work locations as reported by our 19 testing volunteers. Both algorithms perform well, achieving *median* errors of 0.9 miles and 0.83 miles, respectively. We also calculated the distance between the actual home and work locations and the nearest cell tower, finding them to be 0.61 miles (home) and 0.5 miles (work). Moving out to the 95^{th} percentile, the Home algorithm continued to work well, with 3.86 miles of error, whereas the "best-case" algorithm that chooses the cell tower nearest to the user-provided latitude/longitude has 1.12 miles of error. At the 95^{th} percentile error, the Work algorithm's error increases to 21.2 miles, and studying these few cases in more detail revealed that the errors were due to our volunteers not using their cell phone much at work. Given that the majority of our volunteers did use their cell phones at work and given the increasing trend of dropping landlines at both home and work locations, we believe that our Home and Work algorithms are an increasingly useful tool for estimating home and work locations for the general population at large.

5 Example Applications: Commute Distances and Carbon Footprints

In this section, we show how we can apply our algorithms to larger-scale data analysis and policy planning. In one example, we compute home-to-work commute distances for populations of cell phone users aggregated by ZIP code. In another example, we

combine cellular network data with US Census data to estimate the commuting carbon footprints of the same populations.

5.1 Calculating Commute Distances

We define *commute distance* as the geographic distance between a person's home and work locations. Our *HomeWork* algorithm estimates commute distance by calculating the distance between the two locations identified by the Home and Work algorithms described in Section 4.1. Here we evaluate our HomeWork algorithm by comparing its results to those of three other approaches for estimating commute distance: Oracle, TopTwo, and TimeBased.

The Oracle algorithm, given a set of important-location clusters identified by the Important Locations algorithm described in Section 3.2, estimates a volunteer's commute distance as the distance between the two clusters closest to the true home and work locations reported by that volunteer. The Oracle algorithm represents an upper bound on the accuracy of the HomeWork algorithm and is not realizable. We include the remaining two algorithms, TopTwo and TimeBased, to show HomeWork's accuracy relative to simpler algorithms. TopTwo estimates commute distance as the distance between the two important locations with the largest number of network events. The TimeBased algorithm estimates where a person lives as the cluster with the largest number of events on weekends and weekdays between 7PM and 7AM, and it estimates where a person works as the cluster with the largest number of events on weekdays between 1PM and 5PM. To estimate commute distance, TimeBased then calculates the distance between these home and work clusters.

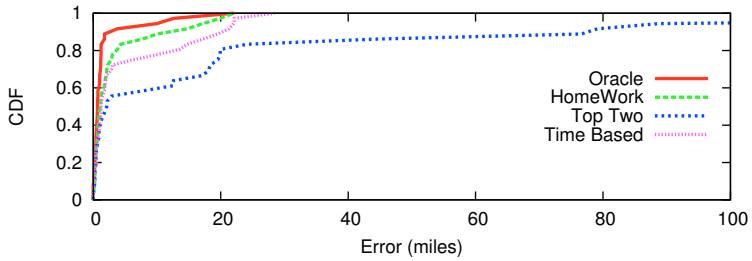


Fig. 5. Error in commute distance for volunteers. The plot is cut at 100 miles for clarity, but the extreme errors for TopTwo extends beyond 200 miles.

Figure 5 plots the CDF of the commute distance error in miles for the Oracle, HomeWork, TopTwo and TimeBased algorithms. The HomeWork algorithm performs very well, estimating the commute distance within 3 miles for 82% of the volunteers. HomeWork not only significantly outperforms both the TopTwo and TimeBased algorithms, which achieve 19.9 miles and 14.2 miles for 82%, respectively, but also is close to Oracle, which achieves 1.23 miles error. The median errors for Oracle, HomeWork, TopTwo and TimeBased are 0.67, 1.16, 2.10 and 1.24 miles, respectively.

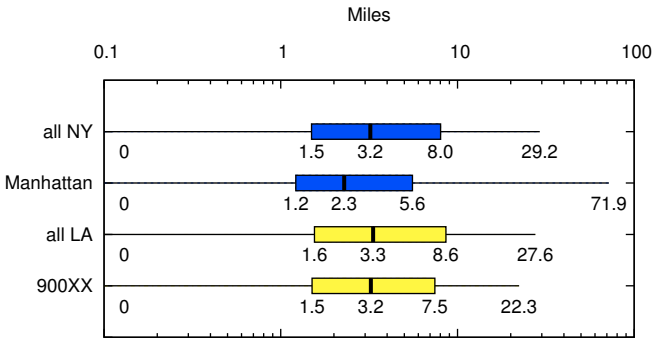


Fig. 6. Box plots of commute distances for Los Angeles and New York.

5.2 Commute Distances in Los Angeles and New York

As an additional check on our work, we compare commute distances as calculated by our HomeWork algorithm to those derived from US Census statistics. In particular, HomeWork estimates the average commute distance for residents of the 891 ZIP codes in our CDR dataset to be 21 and 20 miles for the Los Angeles and New York areas, respectively. Using tables of where people live and work published by the US Bureau of Transportation Statistics [4], we calculate the average commute for residents of the same ZIP codes to be 21 and 19 miles for the Los Angeles and New York areas, respectively. This very close match between HomeWork and census results further validates our approach. It is also important to note that the low cost of our approach makes it practical to regenerate current statistics much more frequently than with a census, for example every few months instead of every ten years.

We now summarize our commute-distance results with the help of boxplots. Boxplots depict five-number summaries of the complete empirical distributions of interest. The "box" represents the 25^{th}, 50^{th}, and 75^{th} percentiles, while the "whiskers" indicate the 5^{th} and 95^{th} percentiles. The horizontal axes show miles on a logarithmic scale. Nearly any difference between our medians is statistically significant due to our large sample sizes.

Figure 6 plots the daily commute distances for Angelenos and New Yorkers. For the greater NY and LA regions, the commute distances are similar, with the median commutes at 3.2 and 3.3 miles, respectively. We also looked in detail at the data from city centers, namely Manhattan (ZIP codes 100xx) and downtown LA (ZIP codes 900xx). We make two observations from the data. First, although the commute distances of city center residents are shorter than that of the general population in both areas, the Manhattanites have a significantly shorter median commute distance compared to all NY (28% smaller) than residents of downtown LA compared to all LA (3% smaller). This is likely because the Los Angeles area is more evenly spread out than NY area. As a result, people who live in the downtown LA commute farther more often than residents of Manhattan. Second, although the median commute distance of Manhattanites is shorter than that of downtown LA residents, 2.3 miles in Manhattan vs. 3.2 miles in downtown LA, the 95^{th} percentile commute distance of Manhattanites is 222% larger at 71.9 miles. These numbers show that when Manhattanites commute far, they commute much

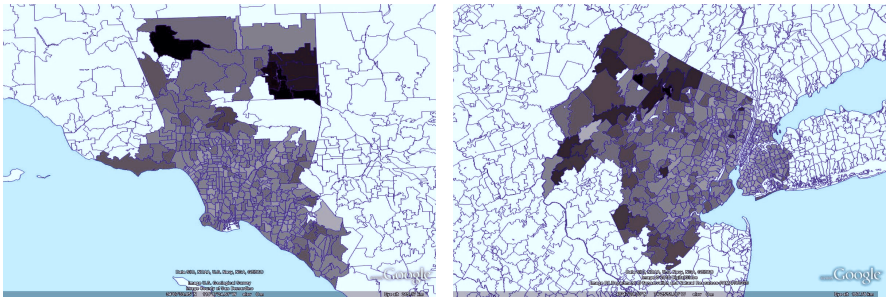
farther than Angelenos. Such long commutes may be due in part to the extensive commuter rail network radiating from Manhattan, which may make such long commutes more feasible.

5.3 Carbon Footprint Estimation

Our final example application makes the extension from commute estimates to carbon footprints. To accurately calculate the carbon footprint of a person's commute, we need to know the length of the commute and the mode of transportation the person uses. Although we can estimate the commute distance of a person using the HomeWork algorithm, the sparsity of our data does not allow us to determine a commuter's mode of transportation. Instead, we determine the mode of transportation of commuters at the ZIP code level using US census data. Specifically, we used Table P30 from the 2000 US census (Summary File 3): "Means of Transportation to Work for Workers 16+ Years." [22] to calculate the percentage of commuters that uses a particular mode of transportation per ZIP code.

The next step is to assign each commuter a mode of transportation that fits her commute pattern best. The intuition behind our approach is that walkers and bikers in each ZIP code are likely to be the people with the shortest commutes. To assign a mode of transportation to each commuter, we first sort the users in each ZIP code according to the length of their commute. If the census reports that P% of commuters in the ZIP code walk or bike to work, then the lowest P% of ranked users in that ZIP code are treated as walkers/bikers with zero carbon emissions. The remaining commuters are assigned the average of the remaining modes of transportation in that ZIP code. For example, assume that we have two commuters in a ZIP code with 50% walkers, 25% drivers, and 25% train passengers. In this example, the commuter with the shorter commute distance is assigned to be a walker and the other user is assigned as driving half the time and taking a train the other half.

Finally, combining this information with the amount of carbon dioxide emitted per person by each mode of transportation [3] allows us to compute the rough amount of carbon dioxide emitted per commuter. Aggregating commuters at the ZIP code level allows us to generate a distribution of carbon dioxide emissions per commuter in each ZIP code.



(a) Los Angeles (b) New York

Fig. 7. Heat maps of median carbon emitted per person for each direction of a commute in the ZIP codes in our study. Darker ZIP codes denote larger carbon footprint. Note that all NY and LA ZIP codes are colored according to the same scale.

Figure 7 shows heat maps of LA and NY, where shading corresponds to the median carbon emission per person in each ZIP code in each direction of a commute. In the New York area, increasing distance from Manhattan correlates with increasing carbon footprint. In contrast, Los Angeles is fairly uniform throughout, with the exception of certain parts of Antelope Valley (in the northeast part of the map), which are separated from downtown LA by a mountain range that must be driven around. These patterns match well with what would be expected from both cities. Popular knowledge indicates that in New York, many people commute into the city center, while in Los Angeles, there is no specific region where people live or work. Manhattan ZIP codes have the lowest carbon footprints of all ZIP codes studied. Specifically, a median amount of carbon dioxide emitted per person is 0.5 kg per trip in Manhattan, 1.07 kg per trip in downtown LA, and 3.7 kg per trip in Antelope Valley.

Generating carbon footprint estimates is a good example of how our technique for computing commuting distances can be combined with already available data to produce new and previously difficult to obtain information.

6 Related Work

There are two broad categories of work closely related to ours. One, there is a body of work that seeks to determine people's important locations based on GPS traces or WiFi beacons. Two, there have been a number of efforts to use cellular network data to find patterns of human mobility. We next survey these two categories of work and contrast them with our own.

Recently, Hightower et al. [11] and then Kim et al. [14] presented algorithms for determining semantically meaningful places based on continuous tracking of GSM and WiFi beacons. Kang et al. [13] explored how clustering locations obtained through WiFi beacons can be used for identifying places people visit. Previous work [1, 16–18] has also explored how semantically meaningful places can be discovered based on series of GPS coordinates. Similarly, Mun et al. [19] estimate the environmental impact of individual travel using GPS traces. Although accurate, these efforts require much finer granularity of data than is available from Call Detail Records. In contrast, we operate on a much larger data set composed of relatively sparse and coarse location samples, which requires a different approach to determining important places in people's lives. In addition, these other approaches collect data using software running on users' devices, which consumes power on those devices and may inhibit large-scale data collection. In contrast, our data is collected by the network for all devices, and does not consume any power on those devices beyond what is consumed by normal use.

There is also a growing body of work attempting to discover patterns of human mobility from cellular network data. González et al. [9] used cellphone records from an unnamed European country to create models of people's movement patterns. Our own recent work [12] characterized the daily range of movement of people in two cities in the United States. Other work has developed algorithms for predicting where a user will travel next [2, 5, 15]. In contrast to the work presented in this paper, that earlier body of work did not seek to attribute importance to any particular location.

Previous attempts to measure the predictability of cell phone users restrict their datasets to highly active users [21] or force phones to provide more frequent location updates than would normally occur [23]. We demonstrate in our work that these steps

are not necessary. We show that we are able to accurately determine important locations across a wide range of usage modes, from highly connected individuals to users that make only a few calls a week.

Finally, Girardin et al. used cell phone usage within cities to determine locations of users in Rome [6], New York City [8], and Florence [7]. They were able to find where people clustered in these cities and the major paths people tended to take. In a sense, they explored the converse of our question. While we ask "to how many locations does this person travel?", they ask "how many people travel to this location?". Furthermore, their work is restricted to a view of a single city, while our work captures travel over a whole country for the subjects in our dataset.

7 Conclusions

This paper has described our work to identify important personal places and movement patterns based on cellular network data. As the central focus of our research, we have proposed and evaluated three algorithms derived from a logistic regression-based analysis of volunteers. The first of these algorithms identifies *Important Places* based on call and text message records. The other two, *Home* and *Work*, narrow down these important places to identify the most likely home and work locations, allowing for the case when they may be one and the same. We validated our algorithms by comparing our results to ground truth from volunteers and to US census data.

Estimating and modeling human mobility is important for many technical and policy reasons. Previously, however, significant challenges have lain in gathering large-scale, comprehensive, and accurate data on which to base such estimates and models. Our work demonstrates that call and text records from cellular networks represent an unobtrusive and accurate way to gather large-scale mobility data. Furthermore, the large degree of aggregation and anonymization allows us to usefully employ this data without unduly impinging on the privacy of any individual.

Viewed broadly, our clustering and location algorithms form a foundation for a range of accurate, low-overhead analyses of human movement and social patterns. As specific examples, this paper demonstrates how we can use home and work identification to perform analyses of commute distances and estimates of commuting carbon footprints. We demonstrate that we can find users' important locations to within 3 miles 88% of the time. We further estimate commute distances within 3 miles of ground truth 82% of the time. In fact, our commute distance estimation errors are quite close to that of an oracle technique, with a median difference from the oracle of only 0.5 miles. Our work is the first to show accurate home and work location estimates and apply them to find carbon emissions from traces that include not just heavy daily cellular phone users, but a nearly universal sample of the user population.

8 Acknowledgments

We thank our shepherd, John Krumm, and the anonymous reviewers for their feedback. Parts of this work were supported by the National Science Foundation under Grant Nos. CNS- 0614949, CNS-0627650, and CNS-0916246. Parts of this work were also supported by a Princeton Engineering fund for Technology for Developing Regions, a research gift from Intel Corporation, and a research internship from AT&T Labs.

References

1. D. Ashbrook and T. Starner. Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 7, 2003.
2. M. A. Bayir, M. Demirbas, and N. Eagle. Discovering spatiotemporal mobility profiles of cellphone users. *World of Wireless, Mobile and Multimedia Networks and Workshops*, 2009.
3. M. J. Bradley and Associates. Comparison of energy use & CO_2 emissions from different transportation modes. Report to American Bus Association, 2007.
4. US Bureau of Transportation Statistics. Downloaded from http://www.transtats.bts.gov.
5. K. Dufková, J.-Y. Le Boudec, L. Kencl, and M. Bjelica. Predicting user-cell association in cellular networks from tracked data. *Intl. workshop on Mobile Entity Localization and Tracking in GPS-less Environnments*, 2009.
6. F. Girardin, F. Calabrese, F. Dal Fiorre, A. Biderman, C. Ratti, and J. Blat. Uncovering the presence and movements of tourists from user-generated content. In *Intn'l Forum on Tourism Statistics*, 2008.
7. F. Girardin, F. Dal Fiore, J. Blat, and C. Ratti. Understanding of tourist dynamics from explicitly disclosed location information. *Symposium on LBS and Telecartography*, 2007.
8. F. Girardin, A. Vaccari, A. Gerber, A. Biderman, and C. Ratti. Towards estimating the presence of visitors from the aggragate mobile phone network activity they generate. In *Intl. Conference on Computers in Urban Planning and Urban Management*, 2009.
9. M. C. González, C. A. Hidalgo, and A.-L. Barabási. Understanding individual human mobility patterns. *Nature*, 453, 2008.
10. J. A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, New York, 1975.
11. J. Hightower, S. Consolvo, A. LaMarca, I. Smith, and J. Hughes. Learning and recognizing the places we go. *Intl. Conference on Ubiquitous Computing*, 2005.
12. S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, J. Rowland, and A. Vasharsvky. A tale of two cities. In *Workshop on Mobile Computing Systems and Applications (HotMobile)*, 2010.
13. J. Kang, W. Welbourne, B. Stewart, and G. Borriello. Extracting places from traces of locations. *Workshop on Wireless Mobile Applications and Services on WLAN Hotspots*, 2004.
14. D. Kim, J. Hightower, R. Govindan, and D. Estrin. Discovering semantically meaningful places from pervasive RF-beacons. *Intl. Conference on Ubiquitous Computing*, 2009.
15. K. Laasonen. *Mining Cell Transition Data*. PhD thesis, University of Helsinki, Finland, 2009.
16. Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma. Mining user similarity based on location history. *Intl. conference on Advances in geographic information systems*, 2008.
17. L. Liao, D. Fox, and H. Kautz. Extracting places and activities from GPS traces using hierarchical conditional random fields. *Intl. Journal of Robotics Research*, 26.
18. N. Marmasse and C. Schmandt. Location-aware information delivery with comMotion. *Intl. Symposium on Handheld and Ubiquitous Computing*, 2000.
19. M. Mun, S. Reddy, K. Shilton, N. Yau, J. Burke, D. Estrin, M. Hansen, E. Howard, R. West, and P. Boda. PEIR, the personal environmental impact report, as a platform for participatory sensing systems research. *Intl. Conference on Mobile Systems, Applications and Services*, 2009.
20. T. Sohn, K. Li, G. Lee, I. Smith, J. Scott, and W. G. Griswold. Place-Its: A study of location-based reminders on mobile phones. *Intl. Conference on Ubiquitous Computing*, 2005.
21. C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327, 2010.
22. US Census Data. Downloaded from http://www.census.gov.
23. H. Zang and J. C. Bolot. Mining call and mobility data to improve paging efficiency in cellular networks. *Intl. conference on Mobile computing and networking*, 2007.