Route Classification Using Cellular Handoff Patterns

Richard A. Becker, Ramon Caceres, Karrie Hanson, Ji Meng Loh, Simon Urbanek,
Alexander Varshavsky and Chris Volinsky
AT&T Labs - Research
180 Park Ave., Florham Park, NJ, USA
rab, ramon, karrie, loh, urbanek, varshavsky, volinsky @research.att.com

ABSTRACT

Understanding utilization of city roads is important for urban planners. In this paper, we show how to use handoff patterns from cellular phone networks to identify which routes people take through a city. Specifically, this paper makes three contributions. First, we show that cellular handoff patterns on a given route are stable across a range of conditions and propose a way to measure stability within and between routes using a variant of Earth Mover's Distance. Second, we present two accurate classification algorithms for matching cellular handoff patterns to routes: one requires test drives on the routes while the other uses signal strength data collected by high-resolution scanners. Finally, we present an application of our algorithms for measuring relative volumes of traffic on routes leading into and out of a specific city, and validate our methods using statistics published by a state transportation authority.

Author Keywords

Handoff patterns, route classification

ACM Classification Keywords

K.8.m Personal Computing: Miscellaneous.

General Terms

Algorithms, Experimentation, Measurement

INTRODUCTION

Urban planners are interested in understanding the mobility patterns of the people who live in and use their cities. This understanding facilitates effective solutions to problems with traffic congestion, parking, vehicular and pedestrian safety, and other aspects of urban living. To gain some knowledge of mobility patterns, planners currently use a combination of census data and vehicle counting. However, the expense of these methods typically results in infrequent data collection and/or small population samples.

Cellular telephone networks have the potential to provide near real-time information about human mobility on a large scale and at a low cost. These networks must know the approximate locations of all affiliated cell phones in order to provide the phones with voice and data services. Since people usually carry their cell phones with them, the location of a phone is a good proxy for the location of its owner.

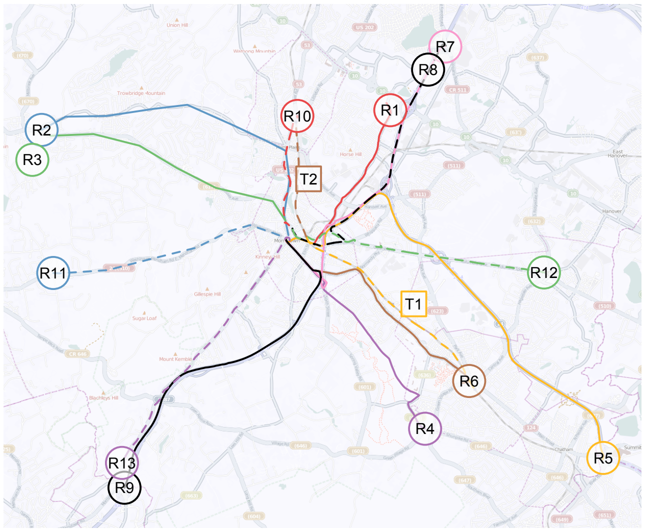

Figure 1. Thirteen driving routes (R1-R13) and two train routes (T1-T2) leading to the center of Anytown. R5, R7, R8, and R9 are freeways; the remainder are other principal commuting routes into town.

This paper explores the use of cellular handoff patterns to identify which routes people take through a city. A handoff pattern is the sequence of cellular antennas that a moving phone uses while engaged in a voice call. However, it is not obvious whether these patterns can be translated into useful route information. The main challenge is location inaccuracy due to the large geographic areas covered by individual antennas, which are often larger than one square mile. Therefore, knowing which antenna a phone is connected to does not immediately reveal what route the phone's owner is traveling on. However, it is possible that knowing sequences of antennas can yield enough information to reveal these routes.

Specifically, we investigate the use of handoff patterns extracted from anonymized Call Detail Records (CDRs). CDRs are collected when a phone is involved in a call, and may contain the full sequence of antennas used by the phone during that call. CDRs are routinely collected by network op-

UbiComp'11, September 17–21, 2011, Beijing, China.
Copyright 2011 ACM 978-1-4503-0630-0/11/09...$10.00.

erators for all active cellular phones, which number in the hundreds of millions in the US and billions worldwide. Furthermore, CDRs are already used for network operation and planning, so additional uses incur little marginal cost. Another advantage of CDRs is that they are generated inside the network and thus do not place any extra burden on the limited resources of mobile phones, particularly their batteries. Despite the fact that our data is restricted to records of phones actively making calls while in transit, the ubiquity and scale of CDR records combined with the likelihood that we will observe a call in a moving vehicle provides us with sufficient data to apply our methods, even in a mid-sized suburban city. We defer the discussion of the limitations of our approach to a later section of the paper. We note, however, that our route identification techniques are independent of how cellular handoff patterns are recorded, whether by the network or by the phones.

Our work explores the following research questions:

1. Are handoff patterns stable across a wide enough range of conditions to be used for identifying routes?
2. Can we devise algorithms that reliably match handoff patterns to routes?
3. Can we derive reliable route utilization statistics from cellular network data?

To answer these questions, we undertook an experimental study of 15 driving and train routes leading into Anytown, a suburban city with roughly 20,000 residents. Figure 1 shows these routes. We used cellular phones to maintain active voice calls while we drove a car and rode the train on these routes. Later, we obtained CDRs from our calls, and used the corresponding handoff patterns to evaluate different route classification techniques. Finally, we applied our best performing techniques to 60 days of anonymized CDRs for all calls handled by one cellular carrier in the Anytown area.

This paper makes the following contributions:

1. We show that cellular handoff patterns are stable across different routes, speeds, directions, phone models, and weather conditions.
2. We propose two algorithms for matching handoff patterns to routes and show that they are accurate. The first uses nearest neighbor classification based on Earth Mover's Distance [8, 9]. The second uses signal strength data to compute the likelihood that a given handoff pattern occurs on a particular route.
3. We show how CDRs, in combination with our algorithms, can be used to determine the relative traffic volumes on roads. We validate these results against statistics published by a state transportation authority.

STABILITY OF HANDOFF PATTERNS

In this section, we show that cellular handoff patterns on a given route are stable over time and across a wide range of conditions. This stability allows us to capture the "typical" handoff pattern for a given route at one time and use it for route classification at another time.

We use the following terminology throughout the rest of this paper. A cell *tower* is a physical structure holding radio antennas and located at a particular latitude and longitude. A *sector* corresponds to a direction from a given cell tower. Each sector is covered by one or more antennas. An *antenna* is a physical device that communicates with mobile phones. Each antenna services a particular cellular technology (e.g., UMTS) and frequency (e.g., 2.1 GHz). Finally, a *handoff pattern* for a call consists of the list of antennas that handled that call, together with the time intervals during which the phone was communicating with each of those antennas. More formally, a handoff pattern for a call handled by n antennas can be expressed as $H = \{(a_i, t_i), i = 1, \ldots, n\}$, where a_i is the identifier of the ith antenna that handled the call and t_i is the duration (in seconds) the call spent on the antenna a_i.

Route Stability Data Collection

To test the stability of handoff patterns on a route, we collected data on a 3-mile stretch of road located in a residential area with many traffic lights and mostly two story buildings. In total, we collected 39 traces under the following varying conditions:

Time Eight months, from September, 2010 to April, 2011.

Phone Model iPhone 3GS, Nokia N95, Samsung Captivate (Android), HTC Aria (Android).

Route Direction North to South and the opposite.

Call Direction Calls originated and received.

Weather Sunny, cloudy, raining, snowing.

Traffic No traffic to heavy traffic.

During each drive, there were one or two pairs of phones in the car, with a call always active between the phones in each pair. In addition, the car's true route was captured using an iPhone application we developed for this purpose. The application recorded the car's location every 10 meters using the iPhone Location API, and uploaded the captured timestamped locations to a server. We later obtained the handoff patterns of all the calls we made during our drives from the network operator's CDRs.

Stability Analysis

Figure 2 shows the geographic map of the area around the 3-mile route. The lines connecting cell sectors to the route represent locations on the route where a call was handed off to the corresponding sector during any of our 39 drives along this route. Portions of the route that do not show handoffs represent areas consistently covered by a single sector.

Figure 3 plots the handoff patterns of the same 39 drives in a way that allows side-by-side comparison. Each horizontal strip represents one drive along the route. Each color represents a different cell sector, using the colors from Figure 2. The drives are sorted by time, with the white strip separating

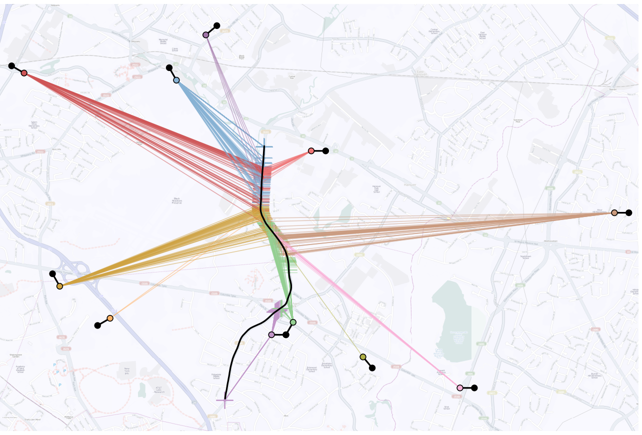

Figure 2. Handoff Stability. The black line marks the 2.8-mile route we drove repetitively to monitor the stability of handoffs under varying conditions. Black dots show the location of cell towers; these are connected to open circles showing the angle of an associated sector. Each thin colored line extends from a point on the route where a handoff occurred to the sector that received the call.

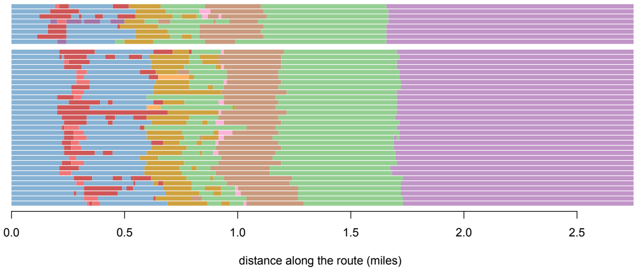

Figure 3. Side-by-side route coverage chart. Each horizontal strip represents one drive along the route. Colors represent sectors covering that portion of the route. The top section (8 calls) were for drives in one direction; the rest in reverse. Note that the handoffs happen at roughly the same locations along the route, and that the set of sectors covering each drive is almost identical.

drives in the two opposite directions. To directly compare the drives in both directions, we reversed the handoff patterns of drives in one direction.

We expect the geographic area for a handoff between two antennas to be roughly the same, regardless of the direction travelled. Indeed, we see each cell sector "covering" a certain geographic area. However, the locations of handoffs during drives in one direction appear to consistently lag relative to drives in the other direction. We believe this is due to the cellular network's desire to keep the phone connected to the current antenna as long as possible to avoid spurious handoffs. Other visualizations (not included in this paper for space reasons),permute the strips in Figure 3 using different grouping variables, and show that handoff patterns are relatively stable across all conditions. The same sectors cover the same road segments consistently, regardless of the phone model, weather conditions, traffic conditions and time.

Note that Figure 3 plots handoff patterns as a function of travelled distance. If plotted as a function of time, however, handoff patterns on the same route show greater differences because the phone may spend different amounts of time on each sector, due, for example, to varying traffic conditions. A good route classification algorithm must handle these time fluctuations.

ROUTE CLASSIFICATION USING NEAREST NEIGHBORS

The stability of handoff patterns across repeated drives on a given route is only part of the story. To match handoff patterns to routes requires that the handoff patterns of those routes are unique, even if the routes are geographically close. In order to assess our ability to classify handoff patterns into actual routes, we identified 15 common routes into and out of Anytown to form the basis of our experimental study. Each route either originates or terminates in the center of the city. We then developed and evaluated two methods for classifying handoff patterns on these routes. The first method uses handoff patterns collected from test drives as the training data for nearest-neighbor classification algorithms, and the second method uses signal-strength data collected on the routes as training data.

Test Drive Data Collection

We collected data on 13 commuter routes and 2 train routes leading into Anytown, a suburban city with approximately 20,000 residents. These routes represent all major ways to get in and out of the city. The route lengths vary from 3 to 6 miles. Many of the routes either partially overlap (e.g., routes R7 and R8) or lay very close to each other (e.g., routes T2 and R10). Overlapping or nearby routes serve to both stress-test our classification algorithms and to reflect reality. Figure 1 shows the 15 routes.

We travelled each route four times, two in each direction, primarily in the Fall of 2010, with a few fill-in drives and train rides in March of 2011. During each drive, there were two phones of different models in the car, one calling the other. As before, we obtained the handoff patterns of all calls from the network operator's CDRs. In total, we collected $4 \times 2 \times 15 = 120$ handoff patterns.

Distance Metrics and Classification Algorithm

Our classification is done via a nearest-neighbor algorithm. For each route, we split the 8 test drives randomly into equal sized training and test sets. For each instance in the test set, we assign the route label of the nearest instance from the training set. We evaluated four different distance metrics, as this choice is crucial for determining the nearest neighbor.

Common Subset Distances

Distances between two handoff patterns can be defined by measuring how much the two patterns have in common [13]. These distances are based on attributes of antennas in the handoff patterns. The larger the intersection between these sets of attributes, the more similar the handoff patterns. We refer to these distances as *common-subset* distances. We defined three common-subset distances that compare these attributes at different levels of granularity: cell towers, sectors, and antennas. The *Common Antennas* distance between two handoff patterns is the number of antennas that appear in

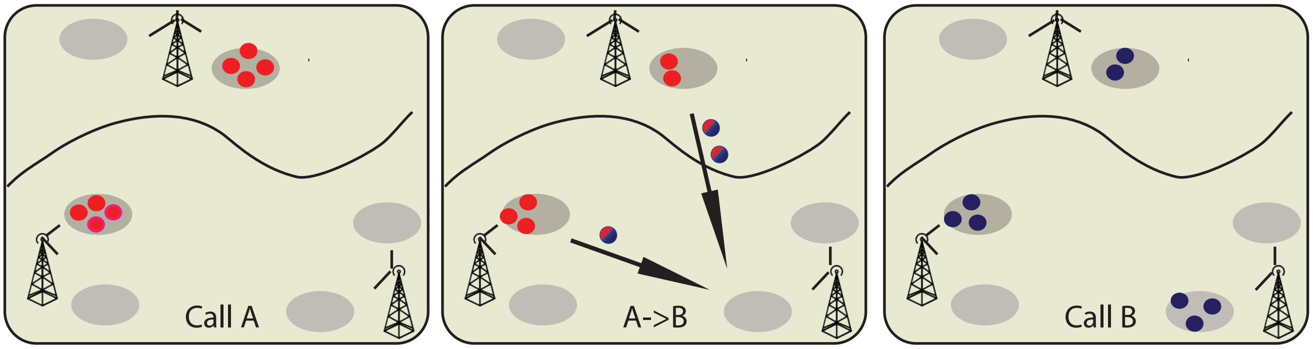

Figure 4. What is the Earth Mover's Distance between call A and call B? Each colored dot represents one minute of connect time to a specific sector (gray ovals). Both calls are 8 minutes long, but distributed differently across sectors. The curved black line represents a road. A->B shows the EMD metric optimally redistributing the minutes from sector to sector, turning call A into call B. The number of minutes moved between each pair of sectors is multiplied by the the Euclidean distance between that pair, summed for all pairs, and divided by total minutes moved, yielding the EMD.

both handoff patterns. Similar definitions apply to *Common Sectors* and *Common Towers* distances.

Classification algorithms based on these common subset distances compute the number of common entities between a handoff pattern in the test set and each of the handoff patterns in the training set. The training route with the highest number of common entities is selected as the best matching route. In the case of a tie, the training handoff pattern with the higher actual number of matching items (as opposed to the common set size) wins.

Earth Mover's Distance

Although common subset distances are good for baselines, they do not account for three important characteristics of the handoff pattern. First, the sequential nature of the handoff pattern is lost, basically reduced to an unordered set of entities to be used in the calculation. Second, temporal information on how long the call spends on each tower is not used by these algorithms. Finally, the cell tower location is not accounted for. Two patterns that differ only by towers that are close to each other should be considered close patterns.

We propose a variant of *Earth Mover's Distance* (EMD) as a distance metric that accounts for all of these characteristics. EMD is traditionally used to measure the differences between images. It was introduced as a technique for image retrieval in the computer vision community, although its roots are far older, [8, 9] and has previously been applied to wireless signals [3]. In the statistical literature, EMD is known as Mallows distance [6] or Wasserstein distance and is used as a distance metric for probability density functions.

Conceptually, imagine a pair of two dimensional images, where each pixel's brightness value is represented by a pile of dirt on that pixel. Pixels with similar brightness have similar amounts of dirt. Now, consider the energy needed to transform one image into the other by moving the piles of dirt. EMD is defined as the minimal energy needed to move the mass of dirt of one image into the locations that result in the target image. One only needs to define how that energy is calculated. Similarly, EMD is defined for arbitrary probability distributions as the mass of probability that needs to be moved to turn one distribution into another,

Figure 4 illustrates a simplified example of how we apply EMD to cellular call data. Given handoff patterns of calls A and B, the figure shows how much and what "work" needs to be performed to convert call A into call B.

More formally, let us define a handoff pattern as a sequence P, with elements $\mathbf{p}_i$ representing the locations of cell sectors in the sequence, and w_{p_i} representing a weight on each of those sectors. Then, let P and Q be two handoff patterns:

$$\begin{aligned} P &= \{(\mathbf{p}_1, w_{p_1}), (\mathbf{p}_2, w_{p_2}), \ldots, (\mathbf{p}_m, w_{p_m})\} && (1)\\ Q &= \{(\mathbf{q}_1, w_{q_1}), (\mathbf{q}_2, w_{q_2}), \ldots, (\mathbf{q}_n, w_{q_n})\}, && (2)\end{aligned}$$

and let $D = [d_{ij}]$ be the geographical ground distance between points $\mathbf{p}_i$ and $\mathbf{q}_j$. Define the flow, f_{ij} as the amount of mass transported between points $\mathbf{p}_i$ and $\mathbf{q}_j$. We want to find $F = [f_{ij}]$, that minimizes the overall cost

$$W(P, Q, F) = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij} \qquad (3)$$

Let $\left[f^*_{ij}\right]$ be the optimal flow that minimizes Equation 3. Then, the Earth Mover's Distance is defined to be

$$EMD(P, Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f^*{}_{ij} d_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f^*{}_{ij}} \qquad (4)$$

In our case, the spatial locations $\mathbf{p}_i$ and $\mathbf{q}_j$ are given by the physical location (latitude and longitude) of sectors. We allow for the directionality of sectors by adjusting this location by a radius r in the direction of the azimuth of the sector. We set the weights w_i to be the duration spent on the particular sectors. We use Euclidean distance

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (\rho t_i - \rho t_j)^2} \qquad (5)$$

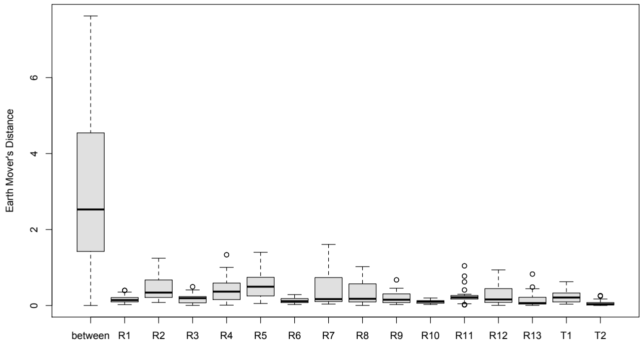

Figure 5. Boxplots show the within-route variation of handoff patterns, as computed by EMD distance. For each route, the within-route distance distribution is much smaller than the between-route distances. This figure shows within-route handoff pattern consistency and that the routes are different from one another and thus can be distinguished.

as our ground distance. In order for d_{ij} to operate on both space and time, we introduce a factor ρ as a conversion factor between the two dimensions to put them on a similar scale. In this way, EMD defines the distance between two distinct calls (handoff patterns), while accounting for temporal and spatial similarity.

One complexity of using EMD for our data is that handoff patterns can have varying length, both in duration and in the sequence of cell sectors, even on the same route (e.g., due to traffic lights). However, even in the extreme case where one sequence is a subset of another, EMD is still an effective distance metric. The imbalance would not contribute to the measure as the excess portion of the longer sequence will not be 'moved' and thus not included in any flow f_{ij}. A by-product of this feature is that the resulting measure is not commutative and hence no longer a proper distance metric. To adjust for this, we use the minimum of EMD(P, Q) and EMD(Q, P) as our distance measure.

Applying EMD in a nearest-neighbor classification task is a straightforward calculation of EMD between the test instances and our training set, then selecting the route label of the nearest one. However, we need one more adjustment to account for situations where the routes of interest are nested, or nearly nested (as in our R2 and R10). In this case, a fraction of a drive along the longer route could be matched well by the shorter route even though the longer route would account for more mass. Hence, for classification, we penalize short matches by adjusting EMD by a multiplicative factor equal to the quotient of the masses $\sum_i w_{p_i} / \sum_j w_{q_j}$, where p_i represents the test sequence and q_i the training sequence.

In this implementation of EMD, both r and ρ are free parameters in our classification algorithm, and so we optimize them using a grid search and cross-validation. The resulting values are $r = 660$m and $\rho = 5$m/s, although the classification is not overly sensitive to the exact values. Our source code is available as the R package `emd` from CRAN.

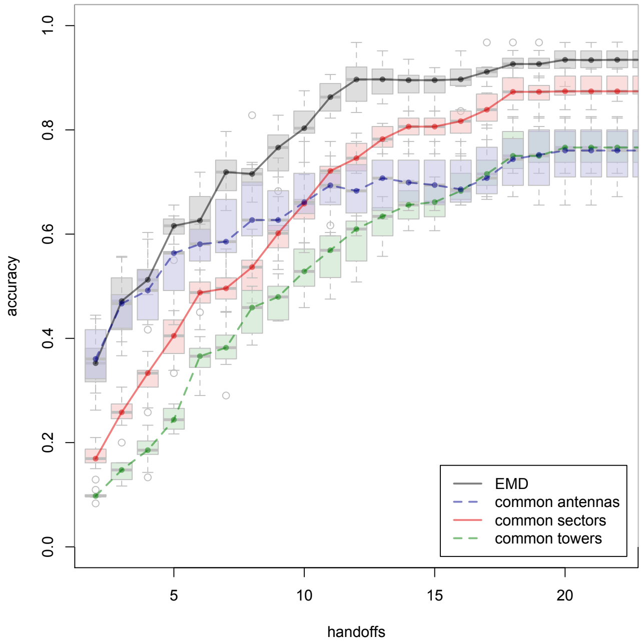

Figure 6. A comparison of nearest-neighbor classification using four distance metrics, with four handoff patterns as training and four as testing per route. Accuracy is shown as a function of the number of handoffs per drive. The boxplots show the range of accuracy over 10 randomly chosen training sets and the colored lines connect the medians of those sets. The EMD metric out-performs the others.

Variability of Handoff Patterns

We assessed the variability of handoff patterns based on the EMD distance metric by calculating all pairwise EMD distances for all of the drives on each of our 15 routes, and compared them to the pairwise distances for drives on different routes. Using the statistical concept of analysis of variance (ANOVA), we expect that the variance among handoff patterns for different drives on the same route to be significantly smaller than the variance between handoff patterns of different routes. Figure 5 shows the distribution of within-route distances plotted as boxplots above each route name, as well as a boxplot for all between-drive pairwise distances.

The figure gives us a good indication of the varying stability of the different drives. It shows that some routes (R1, R6, R10) are extremely stable, with consistently small distances across all pairwise calculations. Other routes are somewhat less stable, showing a wider spread in the boxplot and perhaps a few outliers. However, in all cases the within-route variation is much smaller than the between-routes variation, providing evidence that EMD will be able to differentiate between the routes well.

Performance of the Classification Algorithms

To show that we are effectively able to classify handoff patterns to our defined routes, we split our data into training and test sets, with four randomly selected drives for each route in each set. We fit each test instance to its nearest neighbor in the training set using our four distance metrics, EMD and the three common subset distances, repeating this procedure for 10 different random selections of the training set.

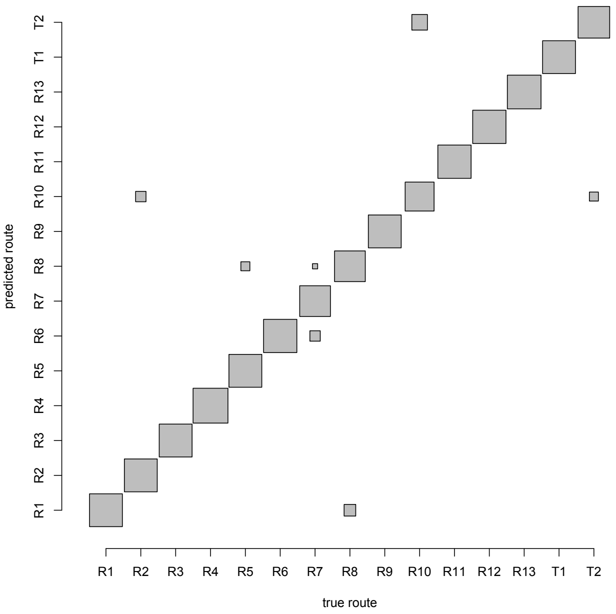

Figure 7. The confusion matrix for EMD algorithm with four training drives shows that most drives were classified correctly. The horizontal axis is the true route and the vertical axis is the predicted route. Correctly identified drives lie on the diagonal; off-diagonal squares are mis-classified. The area of the square is proportional to the number of drives.

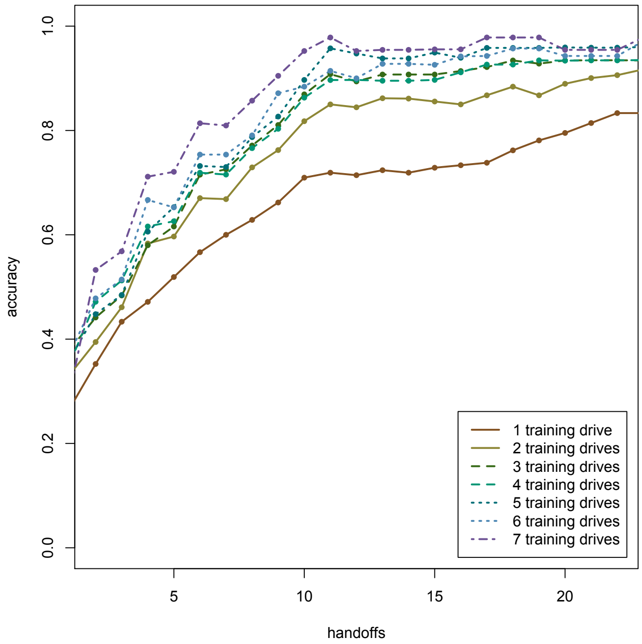

Figure 8. EMD's classification accuracy improves with the number of training samples.

We are also interested in measuring how much information is needed on a single drive in order to classify it correctly. Since all of our routes emerge from the center of Anytown, we assessed the accuracy of our algorithms for a call starting in the center of Anytown as a function of the number of handoffs the algorithm is allowed to observe. Figure 6 shows these results, with boxplots representing the accuracy across all replications for each distance metric and number of handoffs. Colored lines connect medians of the boxplots.

Figure 6 shows that, in general, the prediction accuracy increases as the number of handoffs increases up to a saturation point for all metrics. That is, the farther away the phone moves from the Anytown center, the easier it is to differentiate between routes. The EMD metric performs the best, achieving a median classification accuracy of 90% after 12 handoffs (corresponding roughly to 2 miles). The Common Towers metric performs the worst because it cannot differentiate between handoffs occurring between antennas on the same cell tower. Interestingly, the Common Antennas metric outperforms the Common Sectors metric for up to 10 handoffs, but then performs worse because sometimes phone calls on the same route are handled by different antennas pointing in the same direction.

Figure 7 shows the confusion matrix for the EMD algorithm (the nearest-neighbor classification using the EMD metric) with four training drives using complete test routes. The area of the squares is proportional to the number of drives represented by the square. Squares on the diagonal indicate correctly classified drives, whereas off-diagonal squares are misclassified. Not surprisingly, most of mis-classifications come from overlapping or nearby routes, such as R10 and T2 or R2 and R10 (see Figure 1).

Finally, we studied the effect the number of training drives has on the classification accuracy of the EMD algorithm. Figure 8 shows the performance as we vary the number of training drives from 1 to 7. The results indicate that performance increases with the number of training drives, as the larger training set allows the EMD algorithm to capture more variance on each route. However, we see that with as few as two training drives, performance is quite high, achieving median accuracy of more than 80% after only 12 handoffs. In summary, we showed that the EMD algorithm is well suited to classify handoff patterns under realistic conditions.

ROUTE CLASSIFICATION USING SIGNAL STRENGTHS

The previous section showed that a nearest-neighbor classification algorithm using EMD does a good job matching handoff patterns to routes, requiring as few as two training drives. However, training the algorithm required time-consuming test drives on every target route, and the use of EMD as a distance metric is computationally complex. Although that methodology can be valuable for targeted small deployments where collecting traces of handoff patterns on routes is feasible, a large-scale deployment requires a different approach.

Luckily, cellular network operators routinely use high-resolution scanners to collect GPS-stamped signal-strength measurements from all observable antennas along all major and some minor roads. This process is done for network engineering and maintenance purposes, and it is often referred to as *wardriving* in the research community [5].

In this section, we show that signal-strength data collected by scanners can be used for matching handoff patterns to routes, eliminating the need for training drives. We also present a novel classification algorithm that uses the signal strength data for training and compare its performance with that of the nearest-neighbor classification using EMD.

Scanner Data

We obtained access to scanner traces collected around the Anytown area primarily on September 3, 2010. The traces include GPS-stamped measurements (one per second) along most of our routes. Each measurement contains a list of antennas that were observed at a given location along with the Signal-to-Noise Ratios (SNR), expressed logarithmically in decibels (dB). Unfortunately, we had no scanner data from the train routes.

Although the scanner data did not always correspond exactly to our routes, it did cover the region systematically and we were able to stitch together various parts of the data to match the routes. Since the scanner data is collected at regular intervals throughout the year, we have an opportunity to see how it changes over time, reflecting any changes in network configuration as well as any seasonal differences.

Signal Strength Route Classification Algorithm

The signal strength algorithm classifies a given handoff pattern to one of the routes and it has two stages. In the first stage, it creates a matrix of the maximum SNR values observed on each route from each antenna. If a given antenna has not been observed on a route, the appropriate matrix cell gets a low value of -30dB [1]. We experimented with lower floor values and found no effects on our results. More formally, the algorithm builds a matrix $S = \{(s_{i,j}) : i = 1, \ldots, n; j = 1, \ldots, m\}$, where s_{ij} is the maximum SNR value seen on route i from antenna j.

In the second stage, the algorithm estimates the likelihood of a given handoff pattern appearing on each route by summing up, separately for each route, the maximum SNR values of the antennas appearing in the handoff pattern. The handoff pattern is assigned to the route with the largest likelihood. The intuition is that the antennas with the strongest signals on a route are likely the ones that will appear in the handoff patterns for that route. For example, if a handoff pattern contains three antennas: a_1, a_3, a_5 and there are a total of two routes r_1, r_2, the algorithm computes $S_{r_i} = s_{i,1} + s_{i,3} + s_{i,5}$ for i of 1 and 2, corresponding to the two routes. Summing the SNR logarithms is similar to adding log likelihoods to get the log likelihood of multiple independent events, ignoring any dependence of antennas on one another.

This approach is a weighted variation of the Common Antennas algorithm from the previous section, but the weights and antennas come from the scanned signal-strength data, not from CDRs for our own test drives. Therefore, there is no "training" to be done – the weights take the place of a multitude of test drives.

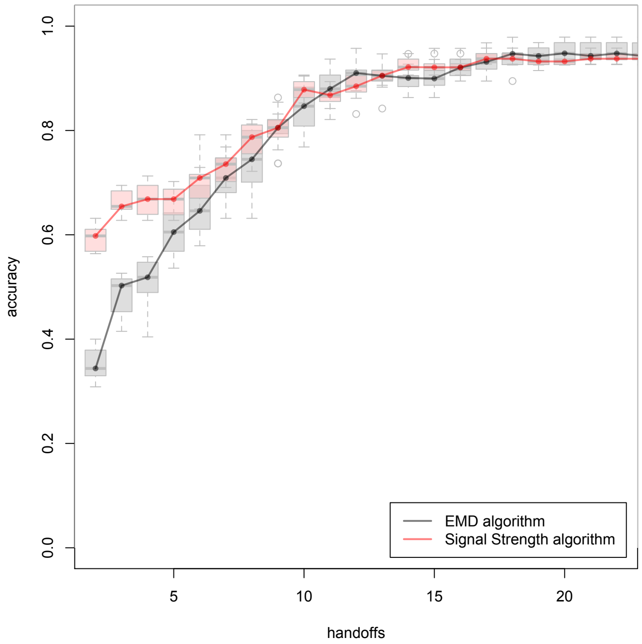

Figure 9. Classification accuracy of the EMD algorithm and the Signal Strength algorithm over 10 random runs.

Figure 9 compares the performance of the Signal Strength and EMD algorithms. Although the Signal Strength algorithm outperforms EMD until up to 8 handoff patterns, the two perform comparably with 9 or more handoff patterns on a drive, achieving more than 90% accuracy. This is a strong result, showing that we can achieve similar accuracy for route classification using two very different approaches.

ESTIMATING RELATIVE TRAFFIC VOLUMES

In this section, we show how our route classification algorithms can be used in practice. Specifically, we show that anonymized CDRs from a network provider, together with our route classification algorithms, can be used to estimate the relative volume of traffic on our Anytown routes.

We had several meetings with urban planners in Anytown, who showed great interest in estimating the flow of traffic along certain routes into and out of the city. When they see congestion in the town center, they do not know where the cars are coming from or headed. And although they can get traffic counts at stationary points from traffic meters, learning about routes entails costly and error-prone surveys of individuals who travel in the area. Detailed route information might help planners to design better traffic signaling and to plan future public transportation infrastructure.

We next describe our CDR collection methodology, the steps we took to protect people's privacy and the limitations of our approach. We then present our estimates for relative traffic volumes on each route. Finally, we compare our estimates to publicly available traffic count data from the Anystate Department of Transportation.

CDR Data Collection

We collected anonymized CDRs from the cellular network of a large US communications service provider. These CDRs capture calls carried by the 35 cell towers located within 5 miles of the center of Anytown, a suburban city with approximately 20,000 residents. These 35 cell towers house approximately 300 antennas pointed in various directions and supporting various radio technologies and frequencies. Our goal was to capture cellular traffic in and around the city and choosing the 5-mile radius allowed us to cover both Anytown proper and its neighboring areas.

In place of the phone number of the person involved in a voice call, each CDR contains an anonymous identifier consisting of a unique integer. Each CDR also contains the starting time and duration of the call, and the locations and azimuths of the cell tower antennas associated with the event. The CDRs contain no information about the second party involved in the call.

We collected voice traffic for 60 days between November 29, 2009, and January 27, 2010, resulting in 15 million voice CDRs for 475,000 unique phones.

Privacy

Given the sensitivity of CDR data, we took several steps to ensure the privacy of individuals. First, only anonymous records were used in this study. The data was collected and anonymized by a party not involved in the data analysis. 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 cell phone 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 antennas associated with a phone during a voice call. The phones were effectively invisible to us outside those times, and we only knew those phone locations at the granularity of an antenna's coverage area, often greater than one square mile.

Limitations

Our CDR data is limited to devices that are actively making a phone call. A US Department of Transportation study [12] estimates that 9% of drivers are using mobile devices while they are driving, but underestimates hands-free usage and omits passenger phone usage. It is possible that this factor differs according to time of day or length of trip, which could skew our results. Nonetheless, combining the 9% usage factor with our carrier's significant market share in the region of the study and the large volume of the overall data provides confidence that our sample sizes are large enough to estimate the traffic flows correctly. This confidence would only increase if we run our algorithms in a market with a larger subscriber base.

Applying our algorithms in areas without the nice hub-and-spoke layouts will require a more careful designation of the roads of interest and is likely to result in different accuracy numbers. However, we believe that our algorithms are applicable to environments of any complexity or size.

Analysis

In the previous sections, we tested our route classification algorithms on handoff patterns known to correspond to one of our selected city routes. In practice, however, it is a challenge to determine whether a voice call was conducted while driving on one of our routes. We used two heuristics to select an appropriate subset of calls. First, we filtered out all CDRs for calls that do not begin or end at the cell tower located at the center of Anytown. Second, we filtered out all CDRs whose handoff patterns include antennas from fewer than 5 distinct cell towers. Although the second step may be too conservative, we wanted to make sure that the remaining CDRs belong to calls made from moving vehicles. After filtering, we were still left with tens of thousands of CDRs.

Figure 10 plots relative traffic volumes as estimated by the EMD algorithm, overlaid on our map of routes into and out of Anytown. The counts are normalized to a count per 1000 cars. The thickness of the line represents the volume estimated on that route. The plot allows comparing the relative number of people who access the town from north and south on the interstate (the black lines) vs. the relative number of people who enter and leave Anytown on secondary state or county roads. The relative traffic volumes as computed by the Signal Strength route classification algorithm are similar and are not included due to space considerations.

Both of our route classification algorithms can deal with calls that are not on any of the target routes. When this happens, the EMD distance will be too large or the signal-strength likelihood too small, and we can set a threshold that indicates that the call does not match. After applying the thresholds on our data, we saw no noticable impact on the relative traffic distributions.

Validation

We present our results as relative volumes instead of absolute volumes since there are many factors that play into whether we will see a particular traveling vehicle: the phone must be active, the user must be a customer of the cellular provider that supplies the data, the phone must use five unique towers, and so on. Because of this, validation of our numbers against readily available government traffic count data is challenging. An additional challenge is that the government traffic data is typically collected using in-street traffic meters or human car counters. Both methods give a count at a static point, while our methods are estimating traffic along a particular route made up of many sequential points.

Nonetheless, we attempted to validate our data with available traffic count information from the Anystate Department of Transportation (DOT). The DOT has a multitude of data available at strategic places around town from traffic meters over the years 2004-2010 [7]. This data allows us to see how

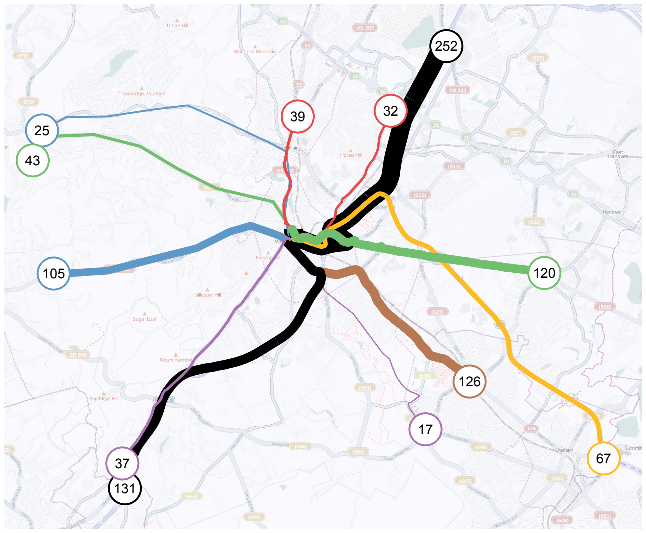

Figure 10. Predicted traffic distribution of 1000 calls. Both the EMD and Signal Strength algorithms were used to estimate traffic volumes on the 13 commuter routes into Anytown. The EMD results are shown as flows into Anytown, with the line width proportional to the estimated volume; the Signal Strength algorithm results are similar.

much daily traffic there is on specific spot on a road. Most of our routes had multiple traffic measurements available at different locations along the route.

Because Anytown is a local hub (it is the largest city in its county), we made the simplifying assumption that all of the traffic on the secondary roads heading into town either originated or terminated at the town center. Since this assumption could not be true for the main interstate running through town, we did not include those routes (R7,R8, and R9) in this analysis. Also, we did not have any traffic counts for any part of R10 that was unique to R10, so we excluded that route as well. Additionally, if a route had multiple measurements of daily traffic at different locations, we selected the minimum of these counts. In most cases, this minimum count was at the furthest point from Anytown, and hence was the best estimate of the number of cars that had travelled the entire length of the route. Finally, we removed a single data point from the DOT data that was highly suspicious as an outlier; it was incongruous with nearby data points in a way that made it appear to be erroneous.

Figure 11 shows a scatterplot of EMD-estimated traffic counts vs. the DOT-supplied traffic counts. The figure shows that the two estimates are closely related, with a correlation coefficient of 0.77. We believe that this result validates our methodology.

To summarize, we showed how CDR data together with our route classification algorithms can be used to estimate relative traffic volumes. We validated this methodology using Anystate DOT traffic counts. The main advantage of our technique vs. simply using Anystate DOT estimates is that we can provide relative traffic estimates much more frequently due to the low cost of our approach. For instance, we could generate figures similar to Figure 10 for different days of the week, times of day, or for special events like a town parade or a holiday.

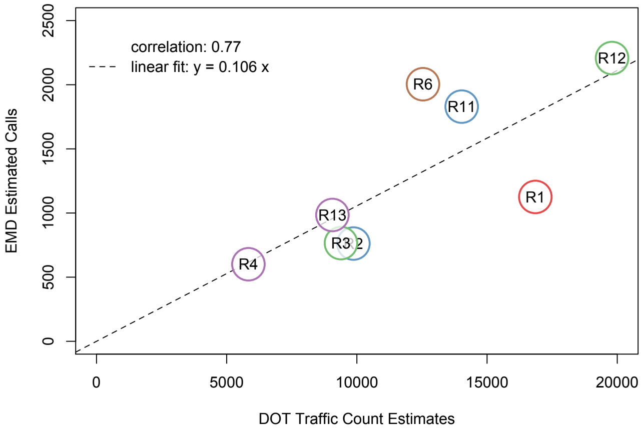

Figure 11. Comparison of traffic volumes estimated by our EMD algorithm against corresponding values obtained from the Anystate Department of Transportation.

RELATED WORK

Mobile phone localization has been an active research topic during the past decade [14]. Placelab [5, 2] was the first system to demonstrate the feasibility of using WiFi and GSM signals for localizing mobile devices. That effort has inspired several follow-up commercial products [15, 16]. Instead of localizing mobile devices, this paper addresses the problem of identifying the route a phone has taken based on cellular handoff patterns.

Krumm et al. [4] developed a Hidden Markov Model (HMM) algorithm for map matching using frequently sampled GPS data. VTrack [11] is a system for travel time estimation using WiFi-based and GPS-based location predictions. VTrack estimates travel times by first mapping location estimates obtained from WiFi and GPS to road segments using a HMM-based algorithm. In contrast to those efforts, we use infrequent changes to the currently associated cellular antenna to match drives to routes.

In a parallel effort to ours, Thiagarajan et al. [10] have developed CTrack, a system for trajectory mapping using cellular base-station fingerprints. Using a two phase HMM-based algorithm and a pre-existing database of location-stamped GSM fingerprints, CTrack is able to match a stream of new GSM fingerprints to road segments with a median accuracy of 75%. CTrack can also utilize information from an accelerometer and a compass to improve its accuracy further. In contrast, our method doesn't require new software to be installed on mobile phones, as we are using the information collected by cellular network operators for billing and maintenance purposes. In addition, our algorithms use only a single cellular antenna at a time, whereas CTrack uses ID and RSSI from up to seven antennas per location. Finally, we analyzed the stability of handoff patterns and showed how to estimate relative traffic volumes on roads.

CONCLUSIONS

CDRs represent perhaps the most abundant records of human mobility available in the world today, and are routinely gathered by cellular network operators for operation and planning. Identifying the trajectories of mobile devices using this CDR data is of great interest to sociologists, civil engineers, and urban planners. In this paper, we showed how to use CDRs to identify which routes people take through a city.

Specifically, we first measured the variability inherent in repeated drives of the same route, and showed that the handoff patterns are relatively stable across different routes, speeds, directions, phone models, and weather conditions. We also employed a novel metric for measuring route variability, based on Earth Mover's Distance, and used it to quantify variability across repeated drives of the same route and between routes.

We then proposed two algorithms for matching handoff patterns to routes and showed their accuracy. The first uses nearest-neighbor classification based on Earth Mover's Distance. The second uses signal strength data to compute the likelihood that a given handoff pattern occurs on a particular route. We showed that the two algorithms have comparable results, with accuracies over 90% for classifying new handoff patterns with only 12 handoffs. Note that although we use handoff patterns extracted from CDRs in our study, our algorithms could be used with cellular handoff patterns captured by the devices themselves. With more and more handsets providing rich programming APIs, the ability to collect this data from the handsets themselves is likely to increase.

Finally, we showed how CDRs, in combination with our algorithms, can be used to estimate the relative traffic volumes on roads, and we validated these estimates against statistics published by a state transportation authority, showing excellent correlation. Currently, such estimates can only be done by laborious placement of in-road or human traffic monitors or by expensive and inaccurate surveys. Conversely, CDRs come in real time and allow studying effects of weather and other disruptions in ways the current sporadic measurements cannot. Our preliminary discussions with urban planners indicate that if proven effective, CDR-based traffic volume estimation could bring significant changes in how they do community planning.

ACKNOWLEDGEMENTS

We thank Mike Orth for providing us with the Nova RF data and Colin Goodall and Siva Prakasam for collecting and anonymizing the CDR data we used in this paper.

REFERENCES

1. A. Buvaneswari, J. M. Graybeal, D. A. James, D. Lambert, C. Liu, and W. M. MacDonald. A statistical view of the transient signals that support a wireless call. *Technometrics*, 49, 2007.
2. M. Y. Chen, T. Sohn, D. Chmelev, D. H. J. Hightower, J. Hughes, A. LaMarca, F. Potter, I. Smith, and A. Varshavsky. Practical metropolitan-scale positioning for gsm phones. In *Proc. of the 8th Int. Conference on Ubiquitous Computing*, Irvine, California, Sept. 2006.
3. K. Kleisouris, B. Firner, R. Howard, Y. Zhang, and R. P. Martin. Detecting intra-room mobility with signal strength descriptors. In *Proc. of the 11th ACM Int. Symposium on Mobile Ad Hoc Networking and Computing*, MobiHoc '10, pages 71–80, 2010.
4. J. Krumm, J. Letchner, and E. Horvitz. Map matching with travel time constraints. In *Society of Automotive Engineers (SAE) 2007 World Congress*, April 2007.
5. A. LaMarca, Y. Chawathe, S. Consolvo, J. Hightower, I. Smith, J. Scott, T. Sohn, J. Howard, J. Hughes, F. Potter, J. Tabert, P. Powledge, G. Borriello, and B. Schilit. Place lab: Device positioning using radio beacons in the wild. In *Proc. of the 3rd Int. Conference on Pervasive Computing*, Lecture Notes in Computer Science. Springer-Verlag, May 2005.
6. E. Levina and P. Bickel. The Earth Mover's Distance is the Mallows Distance: Some Insights from Statistics. In *ICCV 2001*, pages 251–256, 2001.
7. NJ Department of Transportation. Roadway information and traffic counts. www.state.nj.us/transportation/refdata/roadway/traffic.shtm.
8. Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *Proceedings of the Sixth International Conference on Computer Vision*, ICCV '98, pages 59–, Washington, DC, USA, 1998. IEEE Computer Society.
9. Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vision*, 40:99–121, November 2000.
10. A. Thiagarajan, L. S. Ravindranath, H. Balakrishnan, S. Madden, and L. Girod. Accurate, Low-Energy Trajectory Mapping for Mobile Devices. In *8th USENIX Symp. on Networked Systems Design and Implementation (NSDI)*, Boston, MA, March 2011.
11. A. Thiagarajan, L. S. Ravindranath, K. LaCurts, S. Toledo, J. Eriksson, S. Madden, and H. Balakrishnan. VTrack: Accurate, Energy-Aware Traffic Delay Estimation Using Mobile Phones. In *7th ACM Conference on Embedded Networked Sensor Systems (SenSys)*, Berkeley, CA, November 2009.
12. US Department of Transportation. Driver electronic device use in 2009. *Traffic Safety Facts Research Note*, DOT HS 811-372, September 2010.
13. A. Varshavsky, D. Pankratov, J. Krumm, and E. de Lara. Calibree: Calibration-free localization using relative distance estimations. In *Proc. of the 6th International Conference on Pervasive Computing*, May 2008.
14. A. Varshavsky and S. Patel. chapter 7: Location in ubiquitous computing. *Ubiquitous Computing Fundamentals*, 2010.
15. SkyHook Wireless, http://www.skyhookwireless.com.
16. Navizon, http://www.navizon.com.