# Analysis of a Metropolitan-Area Wireless Network

Diane Tang
Stanford University
Gates 3A
Stanford, CA 94305-9030
+1 650 725 1776
dtang@cs.stanford.edu

Mary Baker
Stanford University
Gates 4A
Stanford, CA 94305-9040
+1 650 725 3711
mgbaker@cs.stanford.edu

## ABSTRACT
We analyze a seven-week trace of the Metricom metropolitan-area packet radio wireless network to find how users take advantage of a mobile environment. Such understanding is critical for planning future large-scale mobile network infrastructures. Amongst other results, we find that users typically use the radios during the day and evening. Of the users who move around during the trace (over half), we find that the more locations a user visits on a daily basis, the closer together, on average, those locations are. In addition, the distance users move is a Gaussian distribution around the radius of the network. While these results are only known to be valid for this particular network, we hope future analysis of other networks will add to a growing understanding of mobile network behavior.

## Keywords
Metropolitan-area wireless network.

## 1. INTRODUCTION
Currently, mobile and ad hoc networking is the topic of many research and development efforts. Much of this work focuses on providing future users with network resources and connectivity no matter where they are. Whether the work concentrates on adapting applications to changing user location or on devising new protocols to handle mobility, it is largely based on assumptions of how users will take advantage of a mobile environment. It is difficult to verify these assumptions since we are unaware of any publicly available studies of a sizeable metropolitan-area wireless network. Therefore, many research and development efforts must drive their simulations using assumed models of user movement not derived from observation.

In this paper, we analyze a network trace of the Metricom packet radio network, a metropolitan-area wireless network, to find answers to overall network questions such as when the mobile network is the most active, how active the network gets, where the network is active, as well as radio mobility questions such as how far, how often, and when users move. The answers to such questions are crucial in planning a future mobile network infrastructure, and in understanding how people actually take advantage of a mobile network. While these results are only known to be valid for this particular network, we hope future analysis of other networks will add to a growing understanding of mobile network behavior.

We present several results in this paper, including our finding that the more locations users visit on a daily basis, the closer together, on average, those locations are. In addition, the distance users move is a Gaussian distribution around the radius of the network. We also find that radios are used mostly during the day and evening hours.

In this paper, we first present some background information about the data before we present the actual results from the analysis. The analysis is divided into two parts: overall network behavior and radio mobility.

## 2. BACKGROUND
In this section, we describe the network traced, how the data was collected, and some issues that arose in the data analysis.

### 2.1 Data Collection
The traces we study here were obtained from Metricom [1][5]. Metricom has installed a Ricochet™ packet radio network infrastructure in three major metropolitan areas (San Francisco Bay Area, Washington D.C., and Seattle), as well as in some airports, hotels, and college campuses scattered across the United States. This infrastructure consists of "poletop" repeaters distributed throughout the covered areas. Each poletop is one of two types:

- a wireless repeater, which just forwards packets on to another poletop via the radio interface, or
- a wired access point, which has both a radio interface and a wired interface. Typically, wired access points have 8, 16, or 24 radios on them, as they are the focal point of many other wireless repeaters.

The range of a repeater is normally about half a mile. This range may vary depending on external conditions, such as the weather or the location of buildings and hills.

When a subscriber radio is first turned on, it scans the network for poletops, and chooses one with which to register. This choice is usually based on signal strength, but may also be based on load-balancing considerations. This poletop is responsible for forwarding all packets to and from the Metricom network on behalf of that radio. Radio registrations also occur whenever the radio changes its primary poletop and according to a predetermined pattern when the radio is stationary. Each registration, whether at the same or at a different primary poletop, is logged by a central nameserver.

The trace consists of a nameserver log covering a seven-week period from February 1, 1998 through March 23, 1998, with the exception of three holes in the trace data, during which no registrations were logged: February 16, 6am through 1pm; February 16, 5pm through February 17, 4am; and February 17, 4pm through February 18, noon. The network at the time of the trace consisted of 14,053 poletops and 24,773 radios.

There were a total of 7,726,678 events logged over these seven weeks. Of those, 5,982,846 are registrations and the other 1,743,832 are queries. A registration occurs when the radio informs the nameserver of its current primary poletop, i.e., the poletop to which packets destined for this radio should be sent. A query occurs when the radio queries the nameserver about some other entity in the network, such as another radio or poletop. Queries are usually made at the start of a connection, and the radio can register at different poletops while a connection is on-going.

Each log entry consists of the following information:

- a timestamp taken at the nameserver with accuracy to the second
- a sequence number
- the radio id
- the wired access point used
- the name of the radio's choice for primary poletop
- whether the entry is a registration or a query

A poletop's name is an encoding of its latitude and longitude, so given a poletop's name, we can determine its geographic location and therefore the approximate geographic location of the radio registering with it.

## 2.2 Data Analysis Techniques

In analyzing this trace, the main difficulty is in dealing with the sheer volume of data – 7,726,678 events, 24,773 radios, and 14,053 poletops – making it impractical, if not impossible, to look at each radio by hand. As a result, we needed methods to help gather our results automatically. We turned to a technique commonly used in data mining and machine learning called clustering, also known as an unsupervised learning technique.

Clustering algorithms take a set of points in n-dimensional space and find coherent subsets. Each subset consists of points that are clustered together. The advantages of using clustering algorithms are the ability to categorize radios automatically, and the ability to find groupings that we might not otherwise find. The disadvantages are that the results are dependent on the parameters and distance functions used.

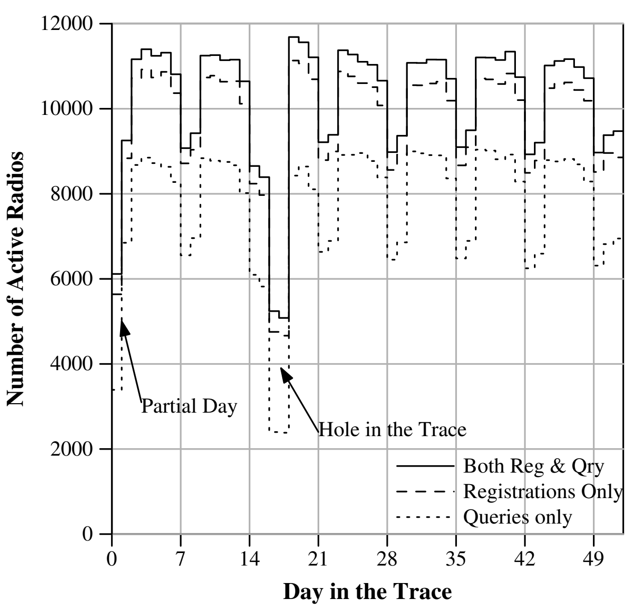



**Figure 1**: *Histogram of the number of active radios on each day of the trace. One corresponds to Sunday, two to Monday, etc. There is a dip around the 16th through the 18th days corresponding to the holes in the trace data.*

We use three different clustering algorithms:

- k-means [9], an iterative clustering algorithm,
- hierarchical agglomerative clustering [7], a tree-forming clustering algorithm,
- and expectation-maximization (EM) [2], a method to find the means and variances of a mixture of Gaussian distributions. EM is especially useful when the range of values differs widely between dimensions.

We use both hierarchical agglomerative and k-means clustering algorithms to help determine when a radio has moved by grouping poletops into clusters, which we also refer to as locations. Without physically moving, the radio may still be within range of up to 20 different poletops at the same time and may therefore register at any of these poletops. We determine that a radio has moved when it registers with a poletop in a different cluster (a different location). In Section 4, we use both EM and k-means to categorize the radios into different patterns of mobility.

# 3. OVERALL NETWORK BEHAVIOR

The first three questions we asked have to do with overall network usage:

1. When is the network the most active?
2. How active does the network get?
3. Where is the network the most active?

The answers to these questions can help network planners plan future extensions to the network infrastructure. Also, they provide a way both to compare a simulated network to an actual network and to provide a basis for simulated networks.

To summarize the results presented in this section, we find that the network is more active on weekdays than on weekends, that the most active poletops handle up to 182 distinct radios an hour and handle over 1,600 registrations an hour, and that the network is most active where there is a high concentration of technical people. We now look at each question in more detail.

To answer the first question, Figure 1 shows that the network is more active on weekdays, especially the days in the middle of the week. This pattern holds regardless of whether we are looking at all events, registrations only, or queries only. Figure 2 shows that the network is least active between 3am and 5am, when most people are asleep. The two lines in Figure 2, one lower in the evening hours and one higher, correspond to weekend and weekdays respectively. Even in Silicon Valley, where the majority of radio users and poletops are, people are less likely to work on weekends. There is a slight decrease in the number of active radios during weekday evenings, perhaps corresponding to when people stop working for the day. There is also a slight rise around 8pm on weekends. We define an active radio as one that has at least one registration or query logged at the nameserver within the given time period.

To find out how active the network gets, for each poletop we first count how many distinct radios register with or query from that poletop over the course of the trace. This ranges from handling only one distinct radio event over the course of the entire trace to handling 6,677 events (6,064 registrations, 434 queries), which occurs at the poletop at Metricom headquarters. Figure 3 shows the distribution within this range.

We also examine how many radios and radio events per hour a poletop needs to handle. Figure 4 shows that while the majority of poletops only handle one radio an hour, some poletops may handle up to 182 distinct radios an hour. Figure 5 shows that most of the time, most poletops handle 500 or fewer radio events per hour, although some particularly busy poletops handle up to 1,600 events in an hour. Note that the shapes of both the dark areas and the peak values in Figure 4 resemble Figure 2 with the low points falling between 3am and 5am, whereas in Figure 5, only the dark areas at the bottom portion of the graph resemble Figure 2. The peaks in Figure 5 do not follow the rise and fall of radio activity. This lack of a clear rise and fall in peak activity could be due to frequent radio registrations regardless of radio activity.

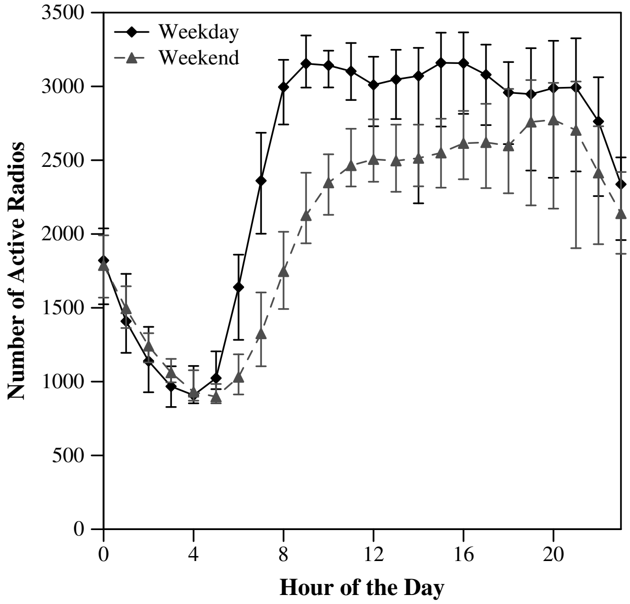



**Figure 2**: *Graph showing the average, minimum, and maximum number of active radios for each hour of the trace. The two sets of lines, one lower in the evening hours and one higher, correspond to the weekend and weekdays respectively. We ignore those days containing the holes in the trace during which no registrations were logged.*

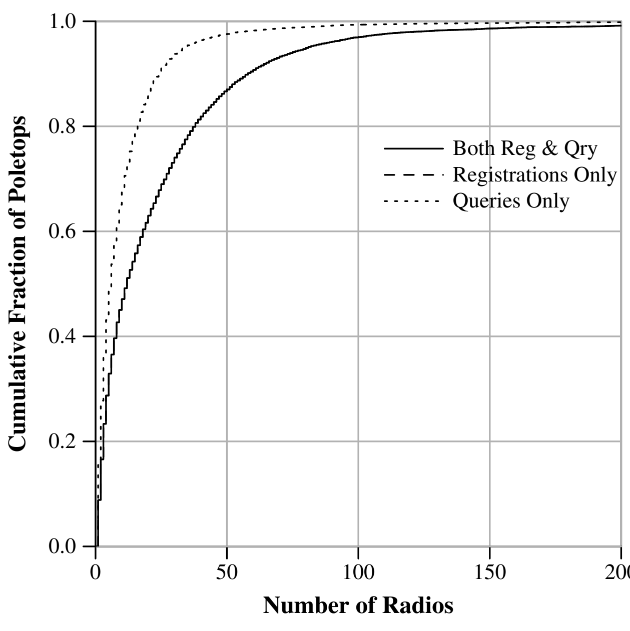



**Figure 3**: *Cumulative Histogram of the number of poletops (y-axis) that handle a certain number of distinct active radios over the course of the entire trace (x-axis). Note that to show the detail, we cut off the tail of the graph, which extends out to 6,677 radios for all events, 6,064 radios for registrations only, and 434 radios for queries only.*

We also look at how often the network needs to handle events, both in the overall network and per poletop. In the overall network, the three largest intervals between two successive events by any radio, discounting the holes in the trace, are 1293 seconds (21 minutes), 744 seconds (12 minutes), and 59 seconds. Each of these intervals occurs only once. Figure 6 shows that by far, however, the network had either no time, or one to two seconds between events. If we look only at registrations in the network, the distribution of intervals is almost identical. However, if we look only at queries, we see that the curve is shallower, with some larger intervals up to 10 to 12 seconds occurring more frequently. This result is due to the many fewer queries than registrations that occur.

Figure 7 shows the cumulative distribution of intervals between events at distinct poletops, rather than in the overall network. This figure differs significantly from Figure 6, with a much longer tail,

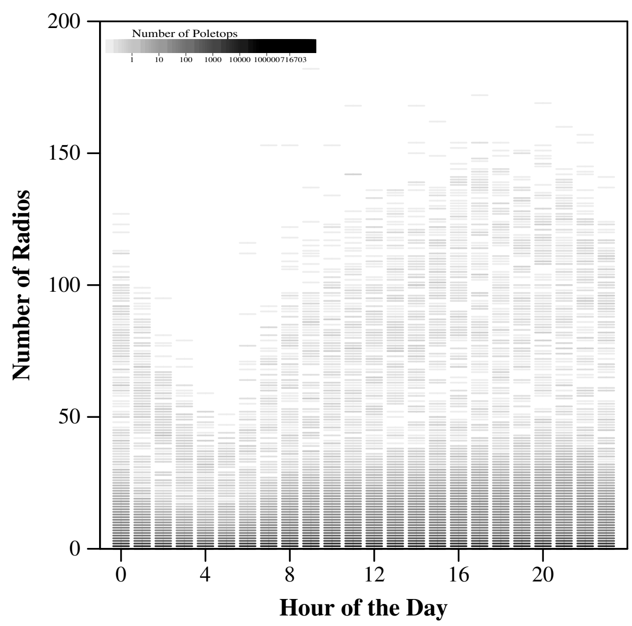



**Figure 4**: *Two-variable Histogram, where the darkness of the bar reflects the number of poletops that receive events from a given number of radios, by hour of the day. No matter what time of day it is, most poletops see events from fewer than 10 radios in an hour. The largest number of radios a poletop receives events from within an hour is 182. Note that the maximum number of radios at any point in the graph is the total number of poletops (14,053) multiplied by the number of times that hour occurs during the trace (once for each of the 52 days).*

due to different poletops having different loads. Some poletops are basically inactive, processing only two or three events during the entire trace, while other poletops handle thousands in an hour. One similarity between the two graphs is the lower curve for queries only, in both cases due to there being fewer queries overall.

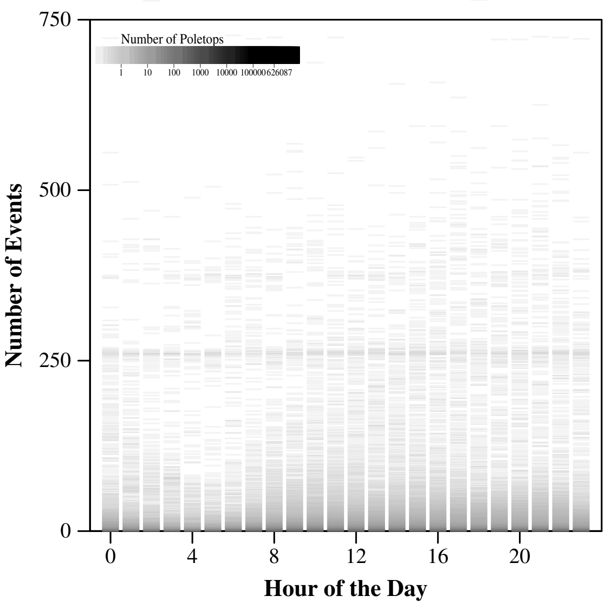



**Figure 5**: *Two-variable Histogram, where the darkness of the bar reflects the number of poletops receiving a given number of events per hour, by hour of the day. Note that the top of the graph, which extends to 1,601 events, has been cut off to show the detail.*

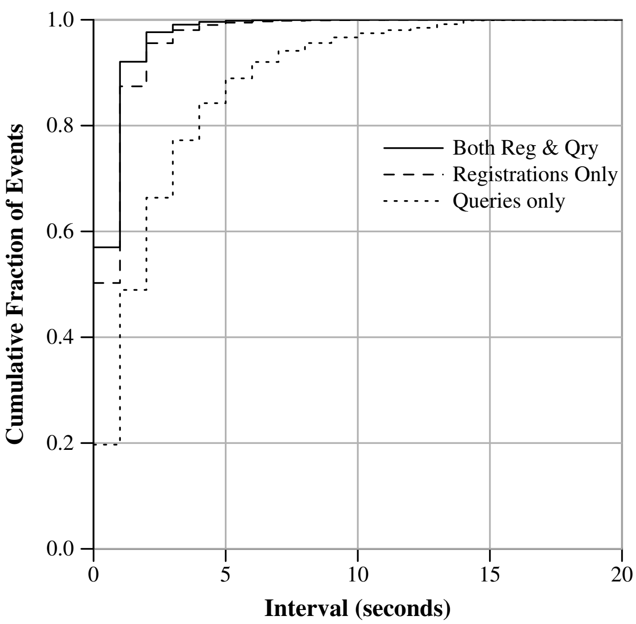



**Figure 6**: *Cumulative Histogram of the interval between successive events over the entire trace as a function of how often that interval occurs. For clarity, we do not plot the tail, which extends to 1,293 seconds.*

Finally, Figure 8 shows the most active poletops in the San Francisco Bay Area. A picture of the entire United States, covering all areas in which Metricom has established an infrastructure is too crowded, so we choose to show just the Bay Area since it has the highest concentration of poletops. The darker dots correspond to the more active poletops. Several hot spots are

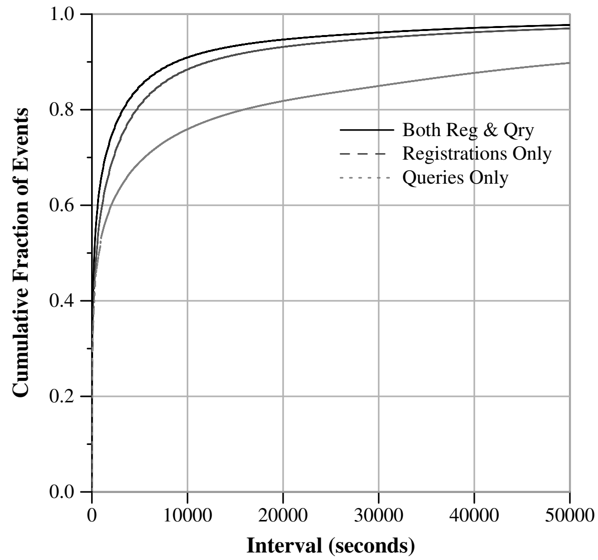



**Figure 7**: *Cumulative Histogram of the interval between successive events at a poletop as a function of how often that interval occurs. For clarity, we do not plot the tail of the graph, which extends out to 4,319,696 seconds (71,994.9 minutes, 1199.9 hours, or 49.9 days) for all events and registrations only, and to 4,370,269 seconds (72,837.8 minutes, 1213.9 hours, or 50.6 days) for queries only.*

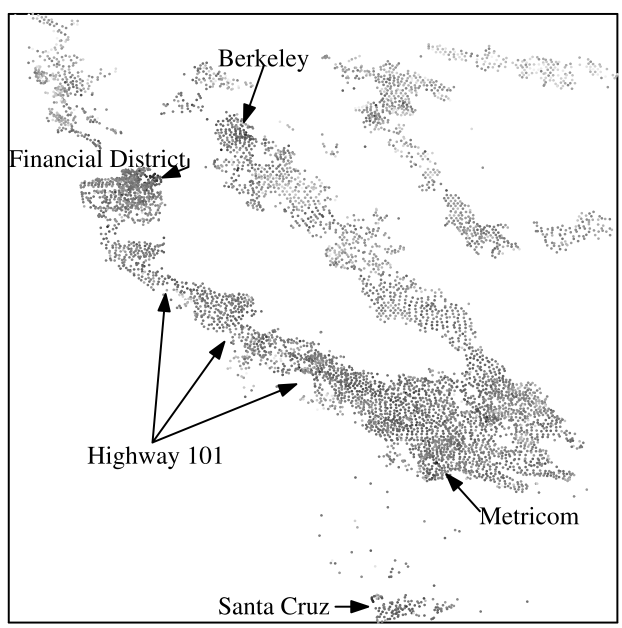



**Figure 8**: *Picture of the San Francisco Bay Area. Each dot corresponds to a poletop. The darker the dot, the more radios visit that poletop over the course of the trace.*

worth noting. First, the dark area in San Francisco corresponds to the Financial District. The second strip of scattered dark dots running northwest-southeast corresponds to Highway 101, on which many high tech companies have their headquarters. Another hot spot is the dark dot in the upper part, which corresponds to Berkeley. Also, towards the bottom is a hot spot corresponding to Metricom headquarters. The lowest island of activity corresponds to Santa Cruz.

We also want to investigate how the activity in the network is distributed geographically, rather than temporally as shown in Figure 6. Now we look at the interval in distance between successive events in the entire network. As we can see in Figure 9, there are several common distances: 100 miles or less, around 650 miles, around 1000 miles, 1400 miles, 2000 miles, and 2400 miles. These distances all correspond either to traffic within a Metricom installation or to the approximate distances between various Metricom installations. Because the Bay Area contains about 65% of the poletops in the Metricom infrastructure, distances of 100 miles or less are prevalent. In other words, two successive registrations in the network are more likely to both be in the Bay Area than to be distributed across the country. However, we can see a Gaussian distribution around each of the common distances. This implies that given the radius of the network, the distance between registrations of any radio in the network is likely to be a Gaussian distribution around the radius.

# 4. RADIO MOBILITY

The other major set of questions we want to answer concerns radio mobility:

1. How often do radios move?
2. How far do they move?
3. Can we identify patterns of mobility?

Answering these questions is crucial for understanding whether and how people actually take advantage of a mobile environment.

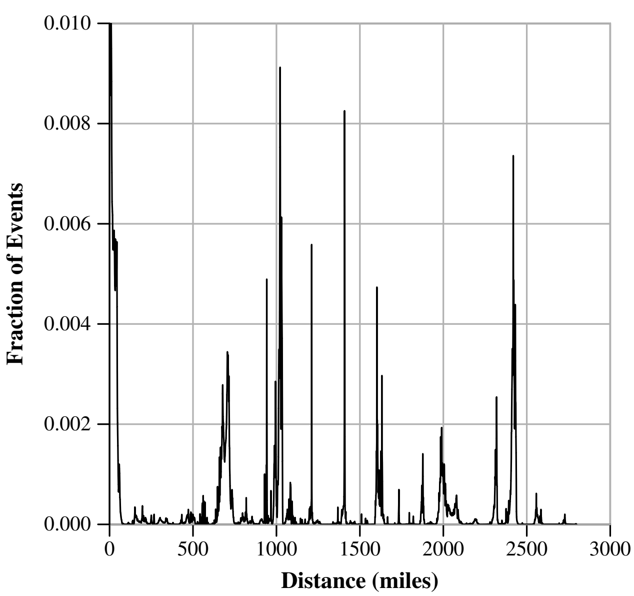



**Figure 9**: *Histogram of the distance between successive events in the network by any radio as a function of how often that distance occurs.*

Also, understanding current radio mobility helps in choosing parameters for simulations of mobile networks. Unrealistic movement models lead to unrealistic simulation results.

To summarize the results in this section, we first find that users who are mobile do not move frequently, and that 64% of all users only appear at one location per day. As for how far users move, most users move within their local area, with fewer users traveling the long distances between different Metricom installations. In addition, as the number of locations visited by a user increases, the average distance traveled between each location decreases. Finally, we are able to find patterns of mobility for users, such as the number of users who are active both day and night versus users active only during the day. We now examine these findings in more detail.

We first calculate the number of different poletops and locations at which radios register over the course of the trace. Figure 10 shows that 42% of all radios are stationary, and that 64% of all radios visit only one location a day. This implies that although users do move around with their laptops, the movement is not very frequent. We can also see the difference between looking at poletops versus locations. While 42% of radios are stationary with respect to location, only 16% of radios are stationary with respect to poletops, meaning that radios often register with poletops that fall in the same cluster or general area. Our location finder underestimates user mobility, so there may be more actual locations than we have found (see Section 5.3).

We also measure how many radios appear, *i.e.*, register or query, at a different location from their last appearance. Figure 11 shows these results. Unsurprisingly, this rate mimics the usage graph in Figure 2, with anywhere from 122 (0.5%) to 1,484 (6.0%) radios changing location at a time. There are, however, some interesting features to the weekday curve in Figure 11. First, there is a dip

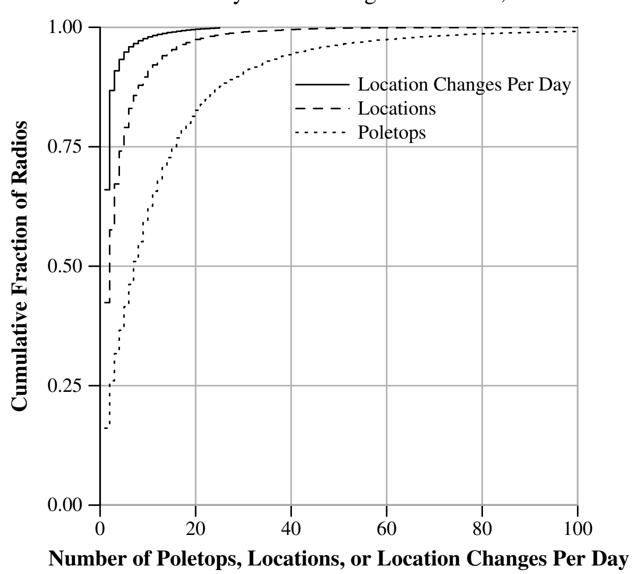



**Figure 10***: Cumulative Histogram of the percentage of radios that visit a given number of poletops, locations, or locations per day, over the course of the trace. For example, 64% of radios visit only one location per day, 42% visit only one location over the course of the entire trace, and 16% of radios visit only one poletop over the course of the entire trace. Note that to show the detail, we cut off the tail of the graph, which extends to 25 locations per day, 176 locations, and 1,025 poletops.*

around 11am to 2pm, approximately corresponding to lunchtime, showing that many people do not take their laptops to lunch wih them. Second, there is an increase in the number of radios changing location around 5pm, about when many people stop working for the day. We still observe the dip in the weekday evenings similar to Figure 2.

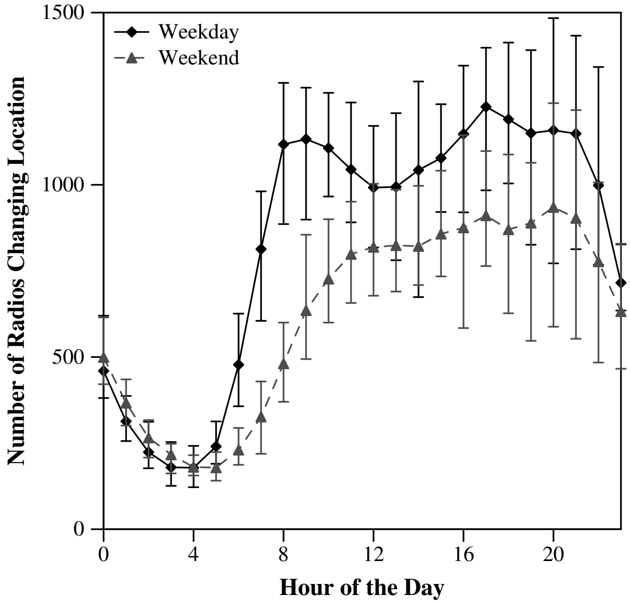



**Figure 11**: *Histogram of the average, minimum, and maximum number of radios changing locations during a particular hour of the day. The time plotted is the second of two successive events.*

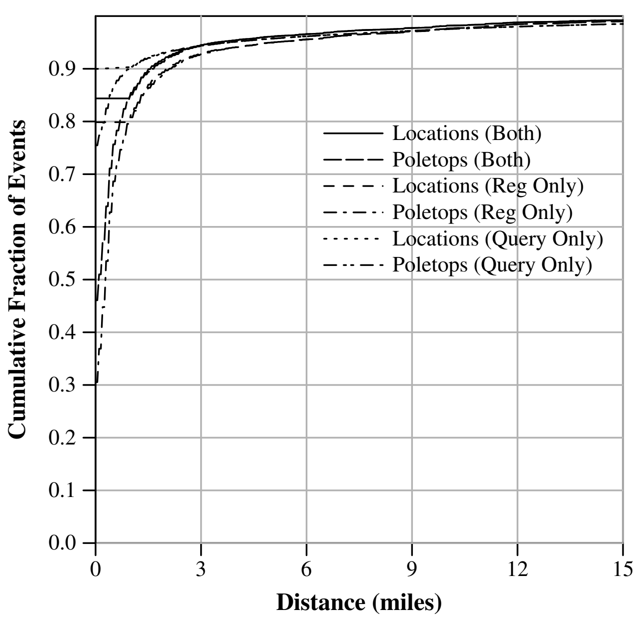



**Figure 12**: *Cumulative Histogram of the number of events as a function of the distance traveled. Note that the tail of the graph (which extends to 2,576 miles) has been removed to show detail. Also note that the two lines, one for poletops, and one for locations, converge at one mile. The straight line between 0 and 0.9 miles for the location graph reflects the fact that locations cluster poletops together; poletops that are close together are considered to be a single location, and therefore represent a movement of distance zero.*

We next look at how far a radio moves when it changes its primary poletop or location. Figure 12 shows the distribution of how often radios move a given distance, in terms of movement both between poletops and between locations. Over 79% of the events (or 80% of registrations, 90% of queries) involve no change in location. If we then look at Figure 13, we can see that while the majority of successive events occur within 50 miles of one another, there are small peaks corresponding to the long-distance routes between airports and major cities in which Metricom has established an infrastructure. While this graph looks similar to Figure 9, cross-country registrations are much less frequent when examining individual user behavior rather than overall network behavior.

We further examine radio movement in Table 1, which shows the breakdown of radios by how often they move, both overall and per day, and how far they move. First, a radio that visits more locations overall does not necessarily also appear at proportionally more locations per day. For instance, the majority of radios that visit three or more locations overall still only appear at one to two locations on a daily basis. Furthermore, only 3.5% of radios that visit more than four locations overall appear at more than four locations per day on average. (Note that one distinct location may be counted as multiple visited locations per day. For example, a user that visits location A followed by location B before returning to location A in one day is counted as appearing at three locations in that day.) This result implies that users who visit many locations do not necessarily move around more on a daily basis, but rather visit the different locations over a longer period of several days or weeks.

More surprisingly, as the number of locations visited on a daily basis increases, the average distance traveled decreases. In other words, the more someone moves around each day, the closer the locations are. This overall decreasing trend may correspond to the flatter distribution of the range of distances among users who visit

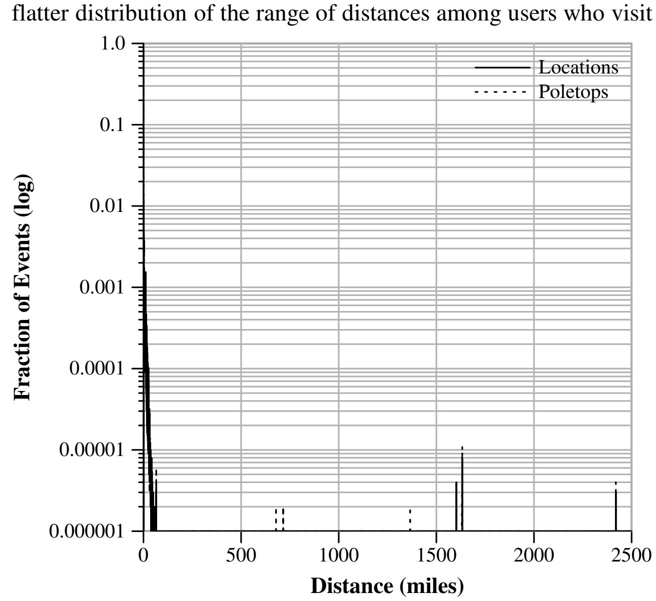



**Figure 13**: *Histogram of the percentage of successive events from a radio as a function of the distance traveled between those events. Note the logarithmic scale on the y-axis.*

| Number of Locations | Number of Radios (%) | Average Distance | Median Distance | Number of Locations per Active Day (max) | Number of Radios (%) | Average Distance | Median Distance |
|---|---|---|---|---|---|---|---|
| 1 | 10,459 (42.2%) | 0 | 0 | 1 (1.0) | 10,459 (100%) | 0 | 0 |
| 2 | 2,746 (15.1%) | 6.9 | 2.9 | 1 (1.0) | 2,152 (84.1%) | 11.6 | 1.6 |
| | | | | 1 – 2 (1.98) | 510 (13.6%) | 2.1 | 1.7 |
| | | | | 2 (2.0) | 84 (2.3%) | 1.7 | 1.3 |
| 3 | 2,371 | 5.2 | 1.6 | 1 (1.0) | 1,367 (57.7%) | 11.3 | 1.7 |
| | | | | 1 – 2 (1.98) | 914 (38.5%) | 3.5 | 1.6 |
| | | | | 2 (2.0) | 47 (2.0%) | 2.3 | 1.9 |
| | | | | 2 – 3 (2.92) | 31 (1.3%) | 2.3 | 2.0 |
| | | | | 3 (3.0) | 12 (0.5%) | 2.2 | 2.2 |
| 4 | 1,694 (6.8%) | 4.2 | 1.5 | 1 (1.0) | 569 (33.6%) | 8.4 | 1.6 |
| | | | | 1 – 2 (1.98) | 1,013 (59.8%) | 4.6 | 1.5 |
| | | | | 2 (2.0) | 44 (2.6%) | 2.3 | 1.5 |
| | | | | 2 – 3 (2.98) | 55 (3.2%) | 2.1 | 1.5 |
| | | | | 3 (3.0) | 3 (0.2%) | 2.0 | 1.9 |
| | | | | 3 – 4 (3.65) | 8 (0.5%) | 1.5 | 1.3 |
| | | | | 4 (4.0) | 2 (0.1%) | 7.9 | 2.7 |
| 5 | 6,503 (26.3%) | 7.9 | 2.4 | 1 (1.0) | 438 (6.7%) | 12.2 | 1.7 |
| | | | | 1 – 2 (1.98) | 4,434 (68.2%) | 11.2 | 1.8 |
| | | | | 2 (2.0) | 126 (1.9%) | 4.2 | 1.3 |
| | | | | 2 – 3 (2.98) | 1,020 (15.8%) | 4.9 | 1.7 |
| | | | | 3 (3.0) | 36 (0.5%) | 6.0 | 1.7 |
| | | | | 3 – 4 (3.96) | 210 (3.2%) | 4.3 | 2.0 |
| | | | | 4 (4.0) | 13 (0.2%) | 4.3 | 2.0 |
| | | | | > 4 (24.48) | 226 (3.5%) | 8.0 | 6.9 |

**Table 1***: How far users move depending on how many locations and locations per day a user visits. For example, 510 users visit only two locations throughout the entire trace, but on average, visit between one and two locations per day.*

multiple locations per day, although the majority of locations are still relatively close together. This distribution could stem, for example, from users using their laptops on a train, such as the train that runs along Highway 101, causing them to "visit" numerous locations on their way to work.

Users who visit one location per day, independent of the total number of locations visited throughout the trace, have a large discrepancy between the average distance traveled and the median distance traveled. This discrepancy is likely the result of these users actually being split into two sub-groups: those who travel locally (for example, people who take their laptops home on the weekends), and those who travel long distances (perhaps people who only use their laptops when travelling). The median distance traveled reflects the usage patterns of this first sub-group, while this second sub-group of users increases the average distance traveled for all the groups of users who, independent of the total number of locations visited throughout the trace, only visit one location per day.

Despite this overall decreasing trend in the average distance, the median distance stays relatively constant, reflecting the many users who only move short distances. Further, the average and median distances, in terms of the total number of locations visited rather than the average number of locations visited per day, are also relatively constant, implying that movement is dependent more on how many locations a user visits on a daily basis than on how many locations a user visits overall.

## 4.1 Radio Queries

Before we look into categorizing radios into different patterns of mobility, we examine radio behavior with respect to queries. Queries are usually made when a radio makes a new connection, although some radios are used as part of Metricom's diagnostic utilities and make many more queries than are normal. Note that a query is made regardless of whether the user is using the radio in modem mode or StarMode (or packet mode) [1]. If the radio is being used in StarMode, the radio could have multiple outstanding connections at a time. Since we do not have information about the end of a connection, we do not account for this situation. Using the radios in modem mode is prevalent, however. Because very few radios are used for diagnostic purposes and very few radios are used in StarMode, we believe that our results are a good approximation.

Our main interest with regard to queries is in how much activity there is between queries. We first look at how many times radios register between queries. Figure 14 shows the cumulative distribution of how many registrations occur before a query, which is indicative of the amount of activity between the beginnings of different connections. The majority of connections occur after 30 or fewer registrations. In Figure 15, we examine how often a radio changes its location or poletop before starting a connection. This graph is indicative of how much movement there is between connections. Note that a connection can be long-lived and last for days, explaining the tail of the graph, which extends out to 1,193 locations or 3,334 poletops. Also note that we count the number of changes in location; if a radio visits location A then B then C then B and then A again, we count this sequence as four location changes. We see in Figure 15 that 70.4% of connections have no movement before them at all, while 95% of radios change location fewer than four times. In other words, most radios do not move around much between connections. Unfortunately, we do not have information about when connections terminate so we do not know how much of the movement occurs while the connection itself is maintained.

## 4.2 Patterns of Mobility

Our final goal is to categorize the radios into different patterns of mobility. One expected pattern, for example, might correspond to a user who moves between using the radio at home and at the office. Another pattern would correspond to the user who just uses the radio at home. Yet another pattern might correspond to a traveling salesman who uses the radio to keep in touch with the warehouse. We are interested in these patterns to see whether and how people take advantage of mobility.

Without comprehensively going through the radios by hand, there is no way to find all the categories, much less categorize each radio, without using clustering algorithms. We use a two-stage clustering because it avoids the difficulty in finding a single stage clustering that balances both the location and time information together. By separating out the clustering into two stages, we can first cluster the radios based on how mobile they are and on some overall frequency parameters. Once we know how mobile the radios are, we can adjust the second stage to reflect the results of the first stage, and sub-cluster based on time of day to determine when radios are active.

More specifically, in the first stage, we cluster the radios based on mobility using nine parameters:

- the total number of unique locations a radio visits throughout the trace,
- the total number of unique locations a radio queries from throughout the trace,
- the average number of locations a radio visits per active day,
- the average number of locations a radio queries from per active day,
- the total number of active days for a radio,
- the total number of query days for a radio,
- the average number of active days in an active week for a radio,
- the average number of query days in an active week for a radio, and
- the average number of query days per week for a radio.

An active day is defined as a day during which the radio visits

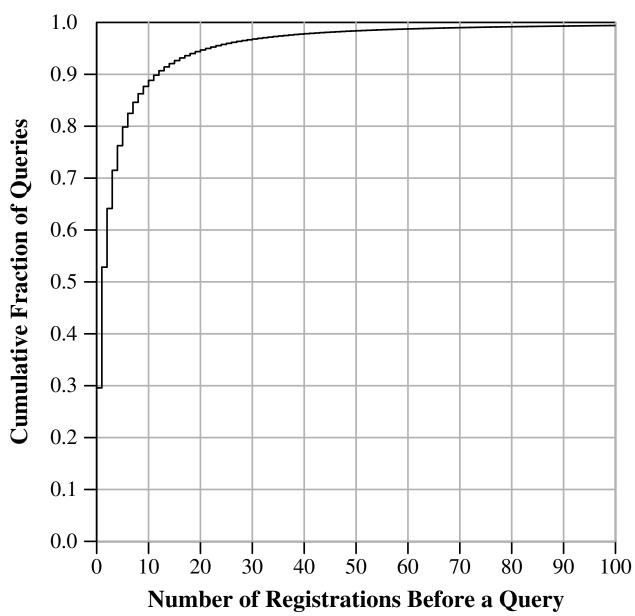



**Figure 14**: *Cumulative Histogram of the number of times a certain number of registrations occur before a query. Note that to show the detail, we cut off the tail of the graph which extends to 7948 registrations between queries.*

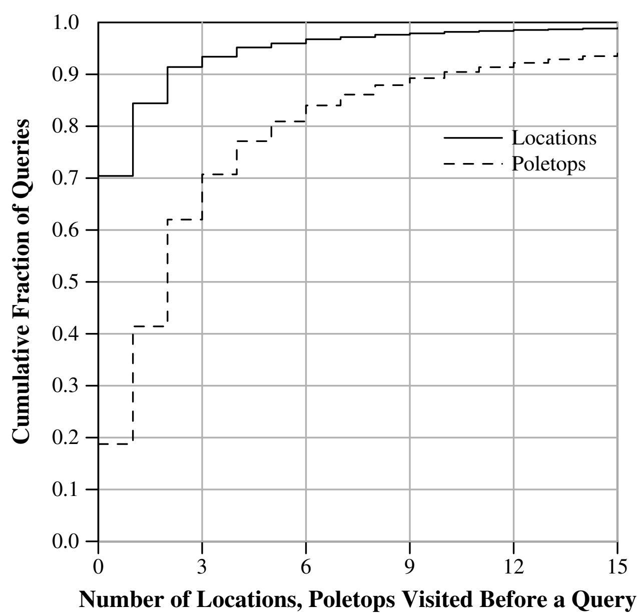



**Figure 15**: *Cumulative Histogram of how often a certain number of poletops or locations are registered at before a query is made. Note that to show the detail, we cut off the tail of the graph which extends to 1,193 locations and 3,334 poletops.*

| Cluster | # of Radios | Total # of Loc's (all, σ, qry only, σ) | Avg # of Loc Changes per Active Day (all, σ, qry only, σ) | Total # of Active Days (all, σ, qry only, σ) | Avg # of active days per active week (all, σ, qry only, σ) | Avg # of queries per day (σ) | Description |
|---|---|---|---|---|---|---|---|
| 1 | 5,877 | 1.02 (0.17)<br>0 (0) | 1 (0)<br>0 (0) | 2.98 (7.34)<br>0 (0) | 1.41 (0.95)<br>0 (0) | 0 (0) | Not used |
| 2 | 4,608 | 1 (0)<br>1 (0) | 1 (0)<br>1 (0) | 27.65 (18.43)<br>20.31 (15.69) | 3.98 (1.94)<br>3.12 (1.68) | 2.41 (2.21) | Stationary (home connectivity?) |
| 3 | 4,765 | 2.98 (1.50)<br>1.60 (0.81) | 1.40 (0.69)<br>1 (0) | 22.66 (17.02)<br>16.04 (13.84) | 3.51 (1.79)<br>2.77 (1.49) | 2.28 (1.51) | Minimal mobility, moderate usage |
| 4 | 1,117 | 5.59 (3.67)<br>2.87 (2.12) | 2.81 (1.66)<br>1.26 (0.77) | 5.63 (3.21)<br>3.78 (2.85) | 1.76 (0.60)<br>1.38 (0.86) | 3.91 (4.32) | Moderate mobility, minimal usage |
| 5 | 1,422 | 5.13 (2.53)<br>3.65 (1.57) | 1.50 (0.41)<br>1.16 (0.10) | 15.50 (5.75)<br>13.63 (5.17) | 2.66 (0.65)<br>2.51 (0.63) | 2.42 (1.05) | Moderate mobility, low usage |
| 6 | 1,603 | 4.55 (2.32)<br>3.17 (1.26) | 1.39 (0.36)<br>1.12 (0.10) | 35.22 (6.88)<br>31.23 (6.26) | 4.69 (0.73)<br>4.24 (0.70) | 2.59 (1.23) | Moderate mobility, moderate usage |
| 7 | 984 | 5.19 (3.0)<br>3.43 (1.66) | 1.84 (0.79)<br>1.25 (0.26) | 47.99 (2.77)<br>46.16 (3.22) | 6.05 (0.33)<br>5.83 (0.39) | 3.64 (2.04) | Moderate mobility, high usage |
| 8 | 882 | 5.23 (3.98)<br>2.41 (1.12) | 9.43 (8.14)<br>1.15 (0.17) | 37.62 (15.36)<br>17.41 (13.19) | 5.15 (1.48)<br>2.76 (1.39) | 1.68 (0.66) | Moderate mobility, train users? |
| 9 | 1,459 | 12.05 (6.44)<br>6.85 (3.46) | 3.06 (1.49)<br>1.43 (0.28) | 22.21 (7.26)<br>19.67 (6.88) | 3.51 (0.78)<br>3.34 (0.81) | 3.23 (1.80) | High mobility, moderate to low usage |
| 10 | 1,299 | 12.14 (7.83)<br>6.83 (4.09) | 4.17 (2.51)<br>1.55 (0.39) | 42.91 (5.63)<br>39.50 (5.88) | 5.51 (0.62)<br>5.13 (0.66) | 3.18 (1.33) | High mobility, high usage |
| 11 | 757 | 18.85 (19.70)<br>9.41 (10.57) | 6.87 (6.02)<br>1.80 (1.31) | 30.92 (15.08)<br>21.34 (16.80) | 4.54 (1.47)<br>3.33 (2.07) | 12.45 (21.83) | High mobility, moderate usage, high connection rate |

**Table 2**: *Results from the first-stage clustering, which results in 11 clusters. Each cluster represents a parameterization for a different pattern of mobility.*

some poletop, a query day is defined as a day during which the radio makes at least one query, and an active week is defined as a week during which at least one day is active. We use EM (expectation maximization) to do this first-stage clustering, since EM takes the differing ranges of the parameters into account, whereas the other clustering algorithms, such as k-means, do not.

We then take each cluster resulting from the first stage, and sub-cluster it based on more specific time information, namely the times of day that the radio is used. Specifically, for each radio we create two histograms counting the number of times the radio has an event per hour of the day, one for weekdays and one for weekends. These histograms show when the radio is active. We then normalize these histograms and use them as the parameters for the second-stage clustering, thus grouping radios with similar active times. We do this sub-clustering, rather than just re-cluster the entire space of radios, so that we can see how each mobility pattern changes by time of day.

We now look at the results from the first-stage clustering and second-stage sub-clustering, shown in Tables 2 and 3, respectively. The first-stage clustering results in eleven clusters. Cluster 1 consists of radios that are essentially not used during the trace. Cluster 2 is comprised of stationary radios, perhaps the ones used on desktop machines for home connectivity. Cluster 3 consists of radios that are minimally mobile, i.e., users that predominantly use their radios from a single location, but may occasionally travel to the office to download some software or data. Clusters 4,5, 6, and 7 all consist of radios that, while they do visit several locations, do not necessarily visit all of those locations during a single day. In other words, the movement is spread out over a larger amount of time. The main difference between these three clusters is in how often the radios are used, with the radios in cluster 4 used the least, and the radios in cluster 6 used the most. However, the radios in cluster 4 also move around the most per day of usage, with more location changes per day. Cluster 8 consists of radios that on any given day, visit approximately twice as many (not necessarily distinct) locations as they see different locations in the entire trace. This behavior could correspond to users who commute by train and use their laptops while on the train, thus visiting most of their locations twice in one day, once on the way to work and once on the way back. Clusters 9 through 11 correspond to radios that move around much more, differing mainly in how often the radios are used and how many connections are formed per day.

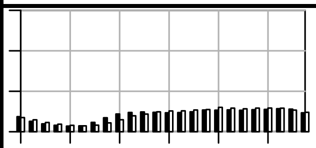
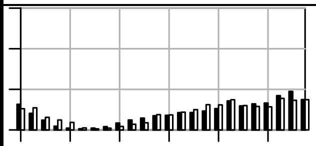
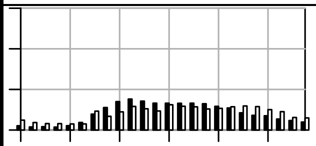
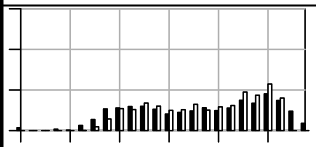
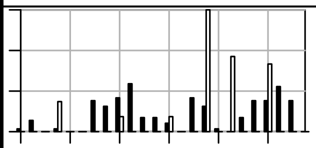
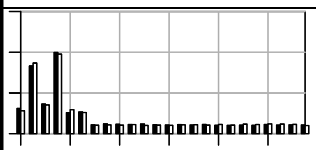
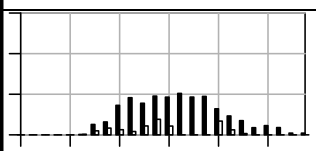
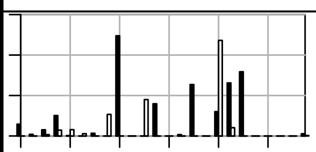
| | Representative Histogram | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | Cluster 7 | Cluster 8 | Cluster 9 | Cluster 10 | Cluster 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 |  | 42.1%<br>1,939 | 26.6%<br>1,267 | | | 12.5%<br>200 | 20.4%<br>201 | 33.9%<br>299 | | 12.5%<br>163 | 28.1%<br>213 |
| 2 |  | | | | | 3.9%<br>62 | 15.3%<br>151 | | | 5.1%<br>66 | |
| 3 |  | 21.4%<br>984 | 39.2%<br>1,865 | 33.1%<br>370 | 46.8%<br>666 | 38.1%<br>611 | 35.9%<br>353 | 49.3%<br>435 | 19.7%<br>287 | 44.3%<br>575 | 37.0%<br>280 |
| 4 |  | 18.9%<br>873 | 8.7%<br>415 | 10.4%<br>116 | 28.6%<br>407 | 42.3%<br>679 | 27.0%<br>266 | 13.5%<br>119 | 71.7%<br>1,046 | 36.9%<br>479 | 28.5%<br>216 |
| 5 |  | 0.2%<br>8 | 1.4%<br>69 | 1.9%<br>21 | 1.9%<br>27 | | | 0.2%<br>2 | 0.2%<br>3 | 0.1%<br>1 | 0.7%<br>5 |
| 6 |  | | | 0.4%<br>4 | | 0.1%<br>2 | 0.9%<br>9 | 0.6%<br>5 | 0.2%<br>3 | 0.8%<br>11 | 0.5%<br>4 |
| 7 |  | 15.2%<br>705 | 23.4%<br>1,114 | 48.6%<br>543 | 20.1%<br>286 | 2.7%<br>43 | | 1.9%<br>17 | 7.9%<br>115 | 0.1%<br>1 | 5.2%<br>39 |
| 8 |  | 2.2%<br>102 | 0.7%<br>35 | 6.1%<br>63 | 2.6%<br>36 | 0.4%<br>6 | 0.4%<br>4 | 0.6%<br>5 | 0.3%<br>5 | 0.2%<br>3 | |

**Table 3**: *Results from the second-stage clustering. Each row corresponds to a time pattern, which is shown with a representative histogram. The x-axis of the histogram is the hour of the day, and the y-axis is the percentage of events occurring during that hour. The black bars correspond to weekday usage, and the white bars correspond to weekend usage. Eight time patterns emerge, and the distribution of these patterns differs amongst the mobility patterns resulting from the first-stage clustering. Note that each first stage cluster corresponds to one column of the table.*

Table 3 shows the eight prevalent timing patterns resulting from the second-stage clustering:

1. a smooth curve similar to Figure 2, with usage balanced across weekends and weekdays.
2. a curve similar to the previous one, but the times are shifted so that the time of least usage is later (around 5 to 6am rather than 4am), perhaps consisting of techies who are used to working later hours.
3. a distribution in which the usage later in the day primarily occurs during the weekend and the usage during the day primarily occurs on weekdays.
4. a distribution in which there is almost no activity all night, and the peaks are a little more obvious. In fact, slight peaks around 9 to 10am and 6 to 8pm can be seen in the weekday and weekend usage.
5. a distribution with very obvious peaks, usually one to two big weekend peaks and two to three smoother

weekday peaks. The histogram pictured is merely representative, since other time usage distributions in this cluster have the peaks at different times.

6. a distribution we call system administration hours, with usage throughout the day, but higher usage late at night.
7. a relatively smooth weekday usage curve, predominantly during the day, with almost no weekend usage.
8. and finally, a very sporadic distribution, with one to two peaks during both the weekday and weekends. Similar to (6), these peaks can be shifted in time.

From Table 3, we see that while we can observe patterns of mobility derived from the first-stage clustering, not all users within a cluster execute their particular mobility pattern at the same time. Instead, users can be further divided into time patterns. For example, Cluster 2 consists of 4,608 stationary radios. This set of radios can further be broken down into users who balance their radio usage across weekends and weekdays (time pattern 1), and users who almost never use their radios at night (time pattern 4, and to a lesser extent, time pattern 3), for example. Also note that not all time patterns occur in every mobility pattern and that the percentages of different time patterns occurring in different mobility patterns vary.

# 5. LESSONS LEARNED

During the course of this analysis, we learned several lessons about the importance of visualization tools, the difficulty in analyzing a trace taken for a purpose other than our own, and the importance of parameter choice in clustering algorithms.

## 5.1 Visualization

The first lesson we learned is that visualization tools are crucial given the volume of data. Without the ability to see a high-level view of the data, it is easy to get bogged down in the details of a very small part of the data set.

In addition, without using different levels of detail, visualization is very difficult, especially given the volume of data involved. In other words, rather than always dealing with the raw registration data, determining intermediate parameters, such as locations and the number of active days, is critical in being able to understand the overall picture.

We used a custom visualization built on top of the Rivet visualization environment [8], combined with our own implementation of the clustering algorithms to help us understand (and debug) the clustering algorithms themselves and to understand the results from the clustering algorithms. Having a custom visualization in which we could implement the exact visualizations and interactions we wanted, adjust the algorithm parameters, and implement and adjust the algorithms themselves was key in understanding the data and the results.

## 5.2 Tracing

The next lesson we learned was about the difficulty in analyzing a trace that was not gathered for the purpose of our analysis, since the trace gathered by Metricom for their own purposes is missing data we would have found useful. Several unanswered questions and ambiguities are a direct result of not having control over what data was recorded and where the data was recorded.

First, one question we wanted to answer was how long a radio stays active, or in other words, how long sessions last. However, session delimiters are not included in the trace. Therefore, session duration results which might be directly calculated from a *tcpdump* trace using TCP SYN and FIN markers, for example, had to be inferred here. In fact, we tried several different methods to infer session duration results, and concluded that while we cannot infer the session duration accurately enough, we can infer that the radio has a linear backoff registration scheme.

Figure 16 shows how often a given interval between successive registrations by a single radio occurs. The histogram shows exponentially decreasing peaks every 600 seconds, or 10 minutes. If the radio is not reset or if the radio does not change its primary poletop, the radio first registers 10 minutes after the initial registration, then 20 minutes, then 30, and so on. Rather than the typical exponential backoff found in networking protocols, a linear backoff is used.

Second, we wanted to investigate patterns of communication between radios. However, the trace has been anonymized to protect Metricom customers' privacy. As a result, we have no information about the other end of a radio's connection, or a mapping from radio to user, and can learn nothing about patterns of communications between radios.

Third, since we have a log from the nameserver only, we could not differentiate problems at the nameserver from problems in the network. For example, the trace includes three periods during which no radio registrations were logged: February 16, 6am through 1pm; February 16, 5pm through February 17 4am; and February 17, 4pm through February 18, noon. We do not know if these holes in the trace were due to a nameserver malfunction or a series of network outages. Metricom has since suggested that this outage is likely a failure in the logging mechanism rather than any

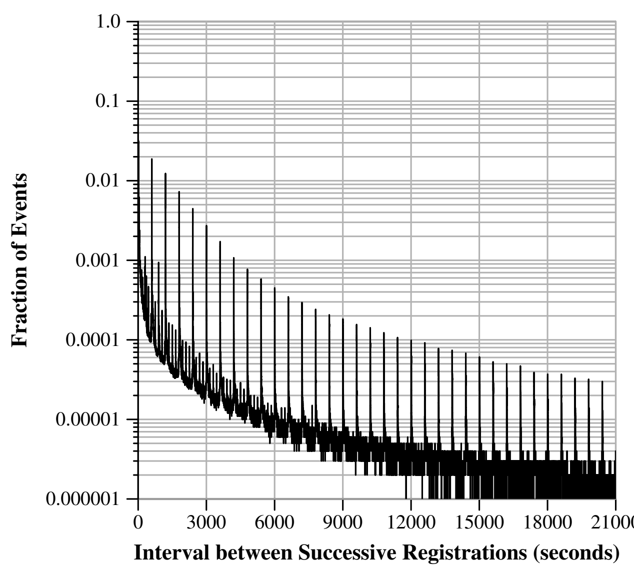



**Figure 16***: Histogram of the number of times a time interval (between successive registrations by a radio) occurs. For example, an interval of length 600 seconds (10 minutes) occurs about 0.025% of the time, or about 150,000 times. Note the logarithmic scale for the y-axis, and that the tail of the graph, which extends to 4,444,118 seconds (74,068.6 minutes, 1234.5 hours, or 51.4 days) has been cut off to show the details.*

network malfunction that would affect users.

Another example of such an ambiguity is a potential network reconfiguration on March 19. On that day, several poletops, previously unseen in the trace, first appear near a poletop at Metricom headquarters and then within the period of an hour, 5pm to 6pm on March 19, reappear at a new location in the Midwest. We do not know whether these poletops malfunctioned, or whether the network was in fact reconfigured. We suspect the poletop radios were being tested before installation in the Midwest.

Finally, the timestamps are the last problem we encountered in analyzing this trace. They are taken at the nameserver, which means that while clock skew is not a problem, network traversal time between the radio and the nameserver is not taken into account. Although we cannot determine the upper bound for this error, we can approximate the average error, since Metricom states that each packet will make no more than three wireless hops before reaching a wired access point. Traversing three wireless hops takes on average between 200ms and 400ms, which is less than the accuracy of the timestamp itself, making this a non-issue.

These unanswered questions and ambiguities are all a direct result of not having the necessary data.

## 5.3 Clustering

The final lesson we learned while doing this analysis is the importance of parameter choice and distance function in both hierarchical and k-means clustering.

For example, when we use hierarchical agglomerative clustering to group poletops into locations, the choice of a cutoff distance is crucial. Hierarchical agglomerative clustering builds a tree by iteratively finding the two closest nodes without parents to become the children of a new parent node. Since the distance associated with each parent node is the distance between its two

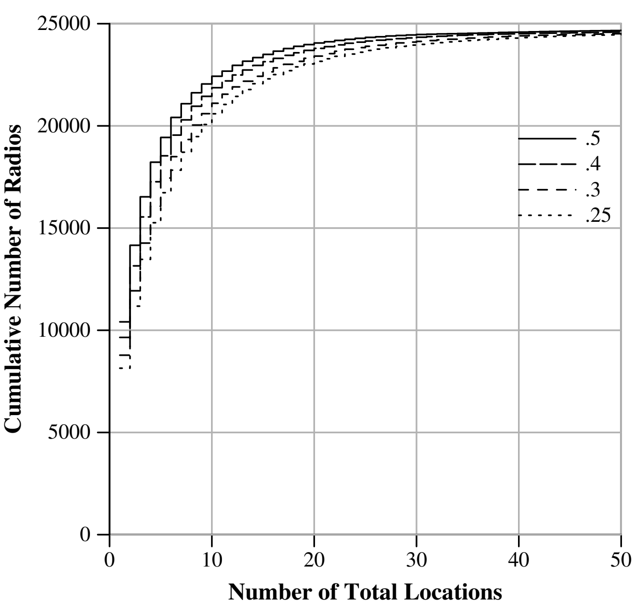



**Figure 17**: *Cumulative Histogram showing how many radios visit a given number of total locations using a minimal distance of 0.25, 0.3, 0.4, or 0.5 miles. The tail of the graph, which extends to 428 locations, has been cut off to show the detail.*

children, we can use this information to differentiate between different clusters. However, we found that there is no good static value to use as a cutoff between clusters.

We use a dynamic cutoff instead. Given all of the distances between children nodes, we look for the first distance which is greater than half a mile. We then look for an exponential increase (i.e., factor of two) over that distance. This new distance is the cutoff used to differentiate clusters of poletops into locations.

We then use k-means to refine the clusters. In k-means clustering, the number of clusters is chosen a priori, and the points are assigned to clusters iteratively. This algorithm is repeated for each radio.

Because we do not know a priori the number of locations a radio visits over the course of the trace, we cannot just use k-means or EM in this situation. Hierarchical agglomerative clustering helps us determine the number of locations for a radio. We also do not use EM in this situation, since the units in both parameters are the same.

The Metricom network already has a coarse granularity: no movement within a building can be detected. We make the network coarser by clustering poletops into locations, with the choice of a minimal distance affecting the resulting coarseness. Figure 17 shows how changing the minimal distance affects the number of locations a radio sees. For example, the most locations any radio visits using 0.5 miles as a minimal distance is 176. Using a minimal distance of 0.25 more than doubles this number to 428. We chose to underestimate the number of locations visited per radio, and thus underestimate the total mobility in the network.

An example of the importance of the choice of distance function in the clustering algorithm occurs in the second-stage clustering used to generate the results in Table 3. Rather than using the standard Euclidean or Manhattan distance when clustering, we modified the standard Euclidean distance function for the clustering based on peak time usage such that the distance between midnight (0) and 11pm (23) is 1 rather than 23.

Finally, the two-stage clustering of radios to find patterns of mobility shows how parameter choice is crucial. When we clustered on poletops merely to find the logical grouping into locations, the choice of parameter was obvious: the latitudinal and longitudinal position of the poletop. For mobility, however, we had many choices in how to parameterize each radio.

When we first clustered based on the overall mobility of the radio and on the frequency of radio usage, some other parameters we could have used were:

1. The total number of registrations, since we might want to differentiate radios that are barely active during the trace. However, this parameter varied so much in value that it influenced the resulting clusters too much.
2. The total number of poletops at which a radio registers. We felt that it would be better to use the location information derived from the poletops, since this is more reflective of the radio's actual movement than the poletops themselves.

We also could have chosen not to use all nine of the parameters we did use. Changing which parameters are used affects the results of the clustering.

Further, when we did the time-based second stage clustering and chose the overall peak times of day for the final dimensionality, we also could have chosen any of the following:

1. Peak time per location. Rather than two overall peaks, base the dimensionality on the total number of locations rather than the number of locations per active day. This choice does not work very well due to users who may visit many more total locations over the course of the trace than they visit per day, leading to unclear results. We could also look at peak time per day of the week, rather than overall peak times.
2. Durations rather than start times. Due to the linear backoff registration scheme, this parameter is hard to quantify.

# 6. RELATED WORK

We know of no other publicly available analysis of a metropolitan mobile network of this scale. However, there are studies of smaller mobile networks.

Researchers at CMU examined their large WaveLAN installation [3]. This study focuses on characterizing how the WaveLAN radio itself behaves, in terms of the error model and signal characteristics given various physical obstacles, rather than on analyzing user behavior in the network.

Another related set of papers is joint work from Berkeley and CMU [6]. The main paper in this set outlines a method for mobile system measurement and evaluation, based on trace modulation rather than on network simulation. This work differs from our own in several ways. First, the parameters they concentrate on deal with latency, bandwidth, and signal strength rather than radio movement. Second, their emphasis is on using these traces to analyze new mobile systems, rather than on understanding the current system. In this paper, our goals are to understand how people use an existing mobile system, with the eventual goal of providing parameters that could be used in simulating mobile networks in the future. Our current focus is on radio movement rather than radio behavior characteristics such as latency and bandwidth.

We previously performed a study of a combined wireless and wired network [4]. However, this study was limited in that only eight users participated rather than the 24,773 in our trace. Also, the Stanford study concentrated more on comparing which end-user applications were used in the wireless versus wired arena, and on determining the characteristics of the wireless network, such as latency and bandwidth.

# 7. CONCLUSION

Although the information in the traces limits the obtainable results and the results themselves are particular to the Metricom network, with its metropolitan-area coverage and high latency (compared to a local-area network), our analysis is a start on understanding how people take advantage of a mobile environment. We find that the more locations users visit on a daily basis, the closer together, on average, those locations are. In addition, the distance users move is a Gaussian distribution around the radius of the network. We also find that radios are used mostly during non-work hours, presumably due to users connecting to a faster network connection during work hours.

Given that many simulations of radio mobility assume a static time between radio movement or some other simple model of user movement, our results present a more realistic model of movement. We hope that these results can be used to guide simulations to yield projected network performance results based on observed movement.

# 8. FUTURE WORK

The main body of future work needed is the comparison of this trace to other mobile network traces. It is possible, even probable, that at least some of the results are a product of the Metricom network and its features, such as its latency, network infrastructure, and trace information or lack thereof. For example, the latency in the Metricom network is high enough that people probably do not use it as they would an in-building wireless network such as WaveLAN. We would like to trace another network ourselves, and see how those results differ from these.

# 9. ACKNOWLEDGMENTS

We thank Metricom, especially Mike Ritter and Loan Nguyen, for providing us with the traces. It is very unusual for a company to share its traces and we greatly appreciate it. Also, we thank Robert Bosch and Chris Stolte for their help with the graphs and the visualization, Lise Getoor and Uri Lerner for their help with the EM algorithm, and Petros Maniatis, Beth Seamans, Guido Appenzeller, Mema Roussopoulos, Ed Swierk, Lucas Pereira, and Andrew Beers for reading paper drafts. This work was supported in part by a generous gift from NTT Mobile Communications Network, Inc. (NTT DoCoMo), a grant from the Okawa Foundation, and a National Physical Science Consortium fellowship.

# 10. REFERENCES

[1] Cheshire, S. and Baker, M. A Wireless Network in MosquitoNet. IEEE Micro, February 1996, 44-52.

[2] Dempster, A.J., Laird, N.M., and Rubin, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. J. Royal Statistical Society, B39:1-38, 1977.

[3] Eckardt, D. and Steenkiste, P. Measurement and Analysis of the Error Characteristics of an In-Building Wireless Network. Computer Communication Review 26, 4 (October 1996), 243-254.

[4] Lai, K., Roussopoulos, M., Tang, D., Zhao, X., and Baker, M. Experiences with a Mobile Testbed. Worldwide Computing and Its Applications, Lectures notes in Computer Science (1368). Berlin: Springer, 1998, 222-237.

[5] Metricom, Inc. http://www.metricom.com.

[6] Noble, B., Satyanarayanan, M., Nguyen, G., and Katz, R. Trace-Based Mobile Network Emulation. Computer Communication Review 27, 4 (October 1997), 51-61.

[7] Rasmussen, E. editors: Frakes, William B. and Baeza-Yates, R. Information Retrieval: Data Structures and Algorithms. Prentice Hall, Upper Saddle River, NJ, 1992, 419-442.

[8] Bosch, R., Stolte, C., Tang, D., Gerth, J., Rosenblum, M., Hanrahan, P. “Rivet: A Flexible Computer Systems Visualization Environment.” Submitted to Computer Graphics and Applications.

[9] Russell, S. and Norvig, P. Artificial Intelligence: A Modern Approach. Prentice-Hall, Upper Saddle River, NJ, 1995.