Clustering Anonymized Mobile Call Detail Records to Find Usage Groups

Richard A. Becker, Ramón Cáceres, Karrie Hanson, Ji Meng Loh, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky

AT&T Labs - Research

1 Introduction

Knowing where and when large numbers of people spend time and how they move between these places can inform many decisions in urban planning and design. Cell phones are an excellent tool for studying human mobility patterns. In the USA, 85% of people 18 and older own a cell phone, 96% of 18-29 year olds own one [1], and most owners willingly carry them wherever they go. They like the feeling of always being in touch, of always having access to other people, and increasingly, having access to the wealth of information on the Internet.

One ubiquitous collection of cell phone usage information is the set of Call Detail Records (CDRs) maintained by cell phone service providers. These CDRs record the time of every voice call or text message exchange (Short Message Service or SMS) along with the approximate location of the cell phone involved. Because they are routinely recorded for resource provisioning and billing, these records provide a comprehensive, inexpensive, and continuing source of information for researchers who want to study human mobility. CDRs are collected from all phones, smart and dumb, and the information is often available quickly, within minutes after the usage events. They provide a complete record of usage, although this means that if the phone is not used, no record is made. Location information is imprecise – using the antenna that handles the call only places the phone in an approximately one square mile region.

We have been analyzing anonymized CDRs to understand what useful information can be gleaned from this source [3, 5, 6]. In this paper, we apply an unsupervised clustering algorithm to CDR data collected in a small city, then investigate the groups of users that result. Members of these groups share the same patterns of cell phone communication, in particular patterns of calling and texting intensity over time. Each group has a specific calling signature, which may be indicative of certain population types such as workers, commuters, and students. Urban planners could use information on how these different groups ebb and flow around the city, in space and time, so as to better understand the urban flow of different sectors of the population. This work focuses on identifying large groups of people based on their shared usage patterns, and is not focused on the individual: throughout this work, we have been careful to preserve individual privacy.

From our analysis, we find seven clusters of users. After ascertaining that the clusters are distinct and plausible, we attempt to learn more about these groups of users. In particular, we look for common locations and times that give the clusters more meaning. For example, we hypothesize that one group consists of commuters calling as they

travel to and from work, based on their voice calling patterns in the hours immediately before and after common working hours. We further show that these people make far more calls when moving than people in the other groups. In a second example, we hypothesize that another group consists of students based on a distinctive usage pattern on weekdays that matches the times of school start, lunch, and dismissal. We then attempt to see if locations of use, represented approximately by antennas, can help confirm the hypothesis. Finally, we note that this cluster is the only one exhibiting that usage pattern.

2 Dataset and Privacy Measures

We collected anonymized CDRs from the cellular network of a large US communications service provider. These CDRs capture transactions carried by the 35 cell towers located within 5 miles of the center of Morristown, NJ, a suburban city with approximately 20,000 residents in the greater New York City metropolitan area. 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 Morristown proper and its neighboring areas.

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

We collected voice and SMS traffic for 60 days between November 29, 2009, and January 27, 2010. In total, we collected 15 million voice CDRs and 26 million SMS CDRs for 475,000 unique phones.

Given the sensitivity of the 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 cellphone user or CDR.

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

Finally, each CDR only included location information for the cellular towers with which a phone was associated during a voice call or at the time of a text message. The phones were effectively invisible to us aside from these events. In addition, we could estimate the phone locations only to the granularity of the coverage area of a cellular antenna. Although the effective radius of an antenna depends upon tower height, radio power and terrain, a given antenna on a cell tower has a location uncertainty of about 1 square mile [10].

3 Method

Clustering is a technique that takes data representing individuals and produces a partitioning of those individuals into groups based on that data. In our case the data for each individual is the relative fraction of all calling made by voice or SMS in each hour of each day of the week. The clustering produces groups where within-group usage patterns are similar and yet the groups are distinct. Unsupervised clustering, which we use here, requires no a-priori knowledge of any individual's cluster membership; the usage data itself determines the clusters.

For each user we encountered, we constructed two vectors that aggregated their voice and SMS usage into 1-hour blocks by day of the week, making each vector $24 \times 7 = 168$ elements long. In order to make voice minutes and SMS counts comparable, we normalized SMS usage counts such that both groups have the same global mean, making one SMS message correspond to 1.5 minutes of voice calling.

Many users in our dataset have very little activity; for example, the median number of SMS events per user per day was 2. Since we are interested in capturing mobility, we decided to perform the clustering on a subset of the users consisting of the most heavy users, namely those with combined activity of at least 10 hours over the 60 days of the study. This subset consists of 26,247 users accounting for 31,169,161 calls/messages. Since our threshold is based on the combined activity, we were careful to verify that we fairly represent top SMS users and top voice users: 71% and 78% of these top users are represented in our sample.

We joined both usage vectors for each user resulting in a vector $2 \times 24 \times 7 = 336$ elements long and normalized them such that each user's vector had the sum of one. We then processed the vectors by k-means clustering [7] using Euclidean distances to produce 7 clusters, as represented in Figure 1.

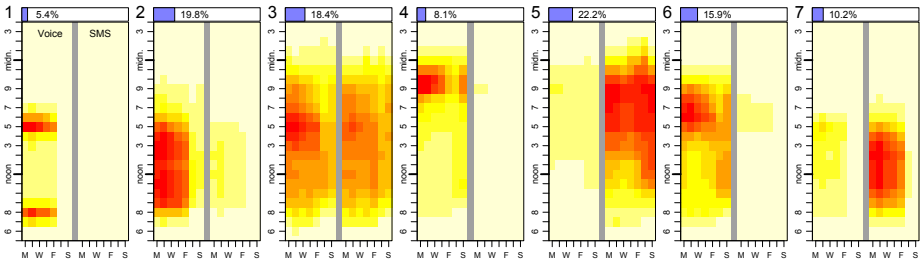
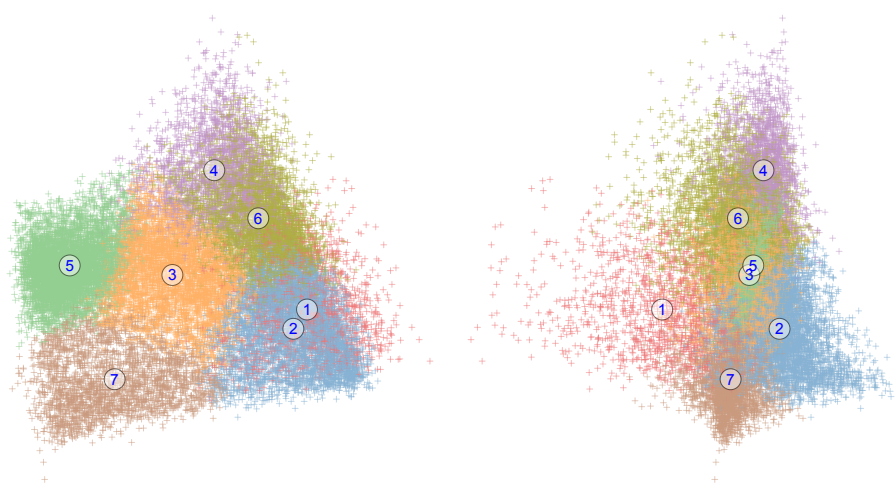


Fig. 1. Seven cellphone usage patterns identified via clustering. Patterns emerge based on voice call and SMS volumes on different days of the week and hours of the day. Voice usage is shown on the left of the gray vertical bars, SMS usage on the right. Darker colors indicate higher volumes. The bar at the top shows the relative size of the cluster, the cluster number is on the top left. For example, cluster 1 shows only voice calls, just before and just after business hours. In contrast, cluster 7 shows primarily SMS usage during business hours.

Because we use Euclidean distance between vectors as our distance metric, there is nothing in the clustering algorithm that enforces coherence between the times or days – the fact that each cluster exhibits patterns with clear hourly and daily groupings shows that these patterns are strongly present in the data.

K-means clustering is an iterative procedure that refines initial estimates of the cluster centroids into final centroids in an attempt to minimize the sums of squares of distances from each point to the nearest cluster centroid. The choice of $k = 7$ clusters was influenced by two factors. First, the curve of within-cluster residual sums of squares as a function of number of clusters starts to tail-off in a region near 5-7 clusters. Second, for 7 or fewer clusters, the sizes of the clusters seem stable across a wide range of starting points for the iterations.

Figure 2 gives a view of the quality of the clustering by showing points for the high-volume users projected into first three principal component axes. Here, each of the 26K points is colored to represent its cluster. The cluster centroids are shown as numbers inside circles, and everything is plotted on the first 3 principal component directions. The 7 clusters are nicely separated in the 3-dimensional space of the first 3 principal components. The first principal component differentiates SMS usage vs. voice usage; the second shows times from early to late; and the third principal component distinguishes between business hours and other times of the day.



(a) 1st (x) and 2nd (y) principal component (b) 3rd (x) and 2nd (y) principal component

Fig. 2. Seven clusters displayed in principal components space for our high volume users. Note that the clusters are well-separated in 3-space. The horizontal axis on plot a) roughly corresponds to SMS vs voice usage; the vertical axis to time of day. Plot b) adds a third principal component that separates clusters 1 and 2, corresponding to business hours and commuting hours.

We produced the clustering based on only the high-volume users. However, any of our 475,00 unique users can be assigned to a cluster by determining the closest cluster centroid to that user. For lower-volume users containing noisy gaps and spikes in usage data it may be advisable to use a metric such as Mallow's distance [8] (also known as Earth Mover's Distance[4]) for cluster assignment. In a real urban planning application, all users with significant usage should be assigned to a cluster in this manner to best understand the cluster mix of the general population. However for descriptive purposes, we now focus on interpretation of the clusters found on the heavy-volume users.

4 Cluster Interpretation

Consider cluster 1. The pattern of usage in Figure 1 shows voice-only usage primarily on weekdays in the hours before and after work. One possible explanation of this is that members of cluster 1 are commuters, calling as they travel prior to and after the workday. To investigate this hypothesis, we attempt to identify the calls made by our high-volume users where there the phone is in transit. How can we do this?

Many of our CDRs identify up to 26 different antennas involved in a voice call. Since we know that even a stationary call can be carried on 3 different cell towers (a tower is a fixed location that typically houses multiple antennas), we use a conservative rule: if a call encounters at least 6 unique towers, it is moving.

Using this definition of a moving call, we computed the fraction of moving calls made by each user in each of our 7 clusters. Cluster 1 has an average value of 11.6%, while the next-higher cluster has an average of 3.2%, which means cluster 1 has more than 3.5 times as many moving calls as any other cluster. The full distribution is shown in Figure 3, where boxplots [9] for each cluster show the 25%, 50%, and 75% quantiles of the moving call distribution. The line in the middle of the box shows the median and the box itself covers the middle 50% of the distribution. Cluster 1's median of 7% is higher than the 75-th percentile of any other cluster. This is compelling evidence to support using the name "commuters" for at least part of cluster 1.

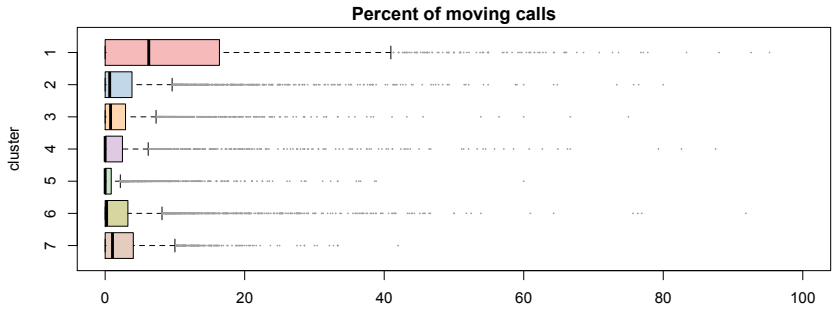


Fig. 3. Boxplots showing the proportion of moving voice calls (those encountering 6 or more unique cell towers) made by each user in clusters 1 to 7. Note that the distribution for cluster 1 is much higher than for the other clusters, leading to the conclusion that cluster 1 has a substantial number of commuters.

Now, let's consider cluster 5. Looking at the bars at the top of Figure 1, we can see that cluster 5 is the largest of the clusters, representing 22% of the 26K clustered users, or about 5700 users. The pattern of use for this cluster is for heavy SMS use between the 3PM and 10PM hours on weekdays, and for beginning earlier and ending later on weekends (about 1AM end on Friday and Saturday). Voice usage is for the same hours, but is much less prevalent than SMS. This sort of pattern is what might be expected for high school and middle school students. (Although there are several colleges in the general vicinity of Morristown, the bulk of our data was collected during winter recess.) Younger people tend to use SMS more than voice [2] and the late start of activity on weekdays seems attuned to school schedules.

How might we confirm the hypothesis that cluster 5 is strongly associated with students? If the group reflects students, we might expect pre- and post-school bursts of usage on school days (students are generally not allowed to use phones during school hours) near the vicinity of school. Figure 4 is an attempt to investigate this hypothesis.

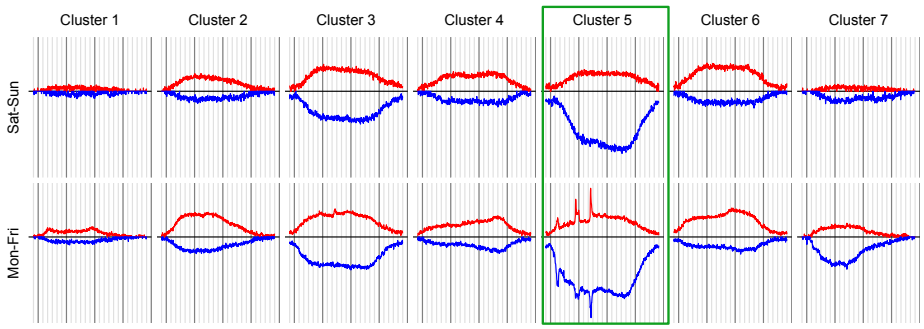


Fig. 4. Lip plot of activity for all clusters on the high school antenna; the top row is for weekends, the bottom row for weekdays. Red lines denote voice usage, blue lines are SMS. The vertical axis is on a square-root scale. The cluster 5 plots are surrounded by a green line and the Cluster 5 weekday plot (the lower plot in the green box) stands out with voice and SMS activity spikes at 7:30AM, 11:30AM and 2:30PM, corresponding to the start of the school day, open lunch, and dismissal.

Figure 4 shows the activities in each of the clusters on the antenna that points in the direction of the city high school. The figure shows a matrix of 'lip plots', where the two rows represent weekends and weekdays, and the columns represent the different clusters. Each plot shows the average volume of voice (red) and SMS (blue, plotted below the y-axis) from 6AM to 6PM during our study period. Cluster 5 stands out because of activity spikes that appear linked to the school day; the spikes occur weekdays at 7:30AM when school begins, at 11:30AM-Noon during an open lunch period, and at 2:30PM dismissal. Voice spikes mirror the SMS spikes, although the SMS usage is greater (the vertical axis is on a square-root scale). Weekend usage, and usage for all other clusters, show smooth changes over time and no similar spikes, indicating that students are not in the other clusters.

We then explore how the usage of cluster 5 is distributed across the different parts of the city. Figure 5 shows cluster 5 usage on all of the different antennas in Morristown. Here the same usage pattern that we see for cluster 5 on the antenna pointing toward the high school (Antenna A3, marked on the plot by a green box) also applies to the antenna pointing at the city's middle school (Antenna E3). An even stronger confirmation of our student hypothesis is that there is no lunch-hour peak for the middle school, because they do not have an open lunch where they are free to leave the school (as at the high school). Also, the middle school seems to have no voice spike, perhaps because the younger students tend to use SMS almost exclusively.

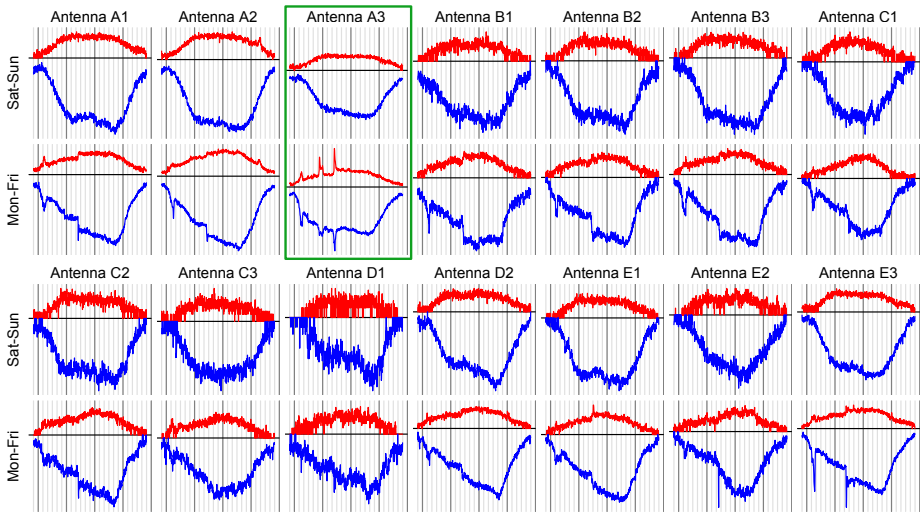


Fig. 5. Lip plots showing cluster 5 activity on all antennas in the city for both Monday-Friday and Saturday-Sunday. The plots with the green outline are the same as those outlined in Figure 4. The strongest spikes occur on the weekdays for the antennas pointing toward the high school (Antenna A3) and middle school (Antenna E3). The two plots for each antenna are scaled together, and the "fuzzy" look of some of the plots is due to fewer calls processed by those antennas, leading to more minute-by-minute variation.

Figure 5 also shows that these distinctive activity spikes are not pronounced at the non-school antennas. However, there are some subtle effects. The pre-school SMS peak appears in lesser amounts across many antennas on the weekdays, but this perhaps may be students texting on their way to school. Similarly, the peak at dismissal seems to correspond to the start of a jump in activity at several of the other antennas, perhaps as students arrive at their post-school destinations and begin their texting in earnest.

As another part of our hypothesis about the student group, we might expect cluster 5 activity to be distributed around the city outside of school hours. This is also evident in Figure 5.

5 Conclusions

Understanding cell phone usage patterns is an important step towards creating applications and services useful for urban communities. Our study involved clustering of usage patterns found in call detail records gathered in a small city. We analyzed anonymized versions of these records, ran a clustering algorithm to find phones that exhibited similar usage patterns, and then looked in more detail at these groups. For a group that we hypothesized were students based on their daily and hourly usage patterns, we were able to find confirmation in more detailed analysis that took into account the antennas that serviced the high school and middle school. The other group we investigated, which we tentatively identified as commuters driving to and from work, indeed turned out to have a much greater proportion of calls involving a cell phone that appeared to be in motion.

References

1. Americans and their gadgets. Technical report, Pew Internet & American Life Project, 2010. http://pewinternet.org/Reports/2010/Gadgets.aspx.
2. Teens and mobile phones: Text messaging explodes as teens embrace it as the centerpiece of their communication strategies with friends. Technical report, Pew Internet & American Life Project, 2010. http://pewinternet.org/Reports/2010/Teens-and-Mobile-Phones.aspx.
3. R. Becker, R. Cáceres, K. Hanson, J. M. Loh, S. Urbanek, A. Vasharsvky, and C. Volinsky. A tale of one city: Using cellular network data for urban planning. Submitted for publication, 2011.
4. F. L. Hitchcock. The distribution of a product from several sources to numerous localities. *Journal of Math. Phys.*, 20:224–230, 1941.
5. S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Vasharsvky. Identifying important places in people's lives from cellular network data. In *Proc. of 9th International Conference on Pervasive Computing (Pervasive)*, 2011. To appear.
6. S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Vasharsvky. Ranges of human mobility in Los Angeles and New York. In *Proc. of 8th International Workshop on Managing Ubiquitous Communications and Services (MUCS)*, 2011.
7. J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. L. Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.
8. C. L. Mallows. A note on asymptotic joint normality. *The Annals of Mathematical Statistics*, 43(2), 1972.
9. R. McGill, J. W. Tukey, and W. A. Larson. Variations of box plots. *The American Statistician*, 32, Feb 1978.
10. C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327, February 2010.