A Tale of Two Cities

Sibren Isaacman[◇], Richard Becker[†], Ramón Cáceres[†],
Stephen Kobourov[⋆], James Rowland[†], Alexander Varshavsky[†]

[◇] Dept. of Electrical Engineering, Princeton University, Princeton, NJ, USA
[†] AT&T Labs – Research, Florham Park, NJ, USA
[⋆] Dept. of Computer Science, University of Arizona, Tucson, AZ, USA

[◇] isaacman@princeton.edu
[†] {rab,ramon,jrr,varshavsky}@research.att.com
[⋆] kobourov@cs.arizona.edu

ABSTRACT

An improved understanding of human mobility patterns would yield insights into a variety of important societal issues such as the environmental impact of daily commutes. Location information from cellular wireless networks constitutes a powerful potential tool for studying these patterns. In this work we use anonymous and aggregate statistics of approximate cell phone locations in Los Angeles and New York City to demonstrate clearly different mobility patterns between the two cities. For example, we show that Angelenos have median daily travel distances two times greater than New Yorkers, but that the most mobile 25% of New Yorkers travel six times farther than their LA counterparts.

1. INTRODUCTION

Characterizing human mobility patterns is key to developing a deeper understanding of the effects of human movement. For example, the impact of human travel on our environment cannot be understood without such a characterization. Similarly, understanding and modeling the ways in which disease spreads around the world hinges on a clear picture of the ways that humans themselves spread[1].

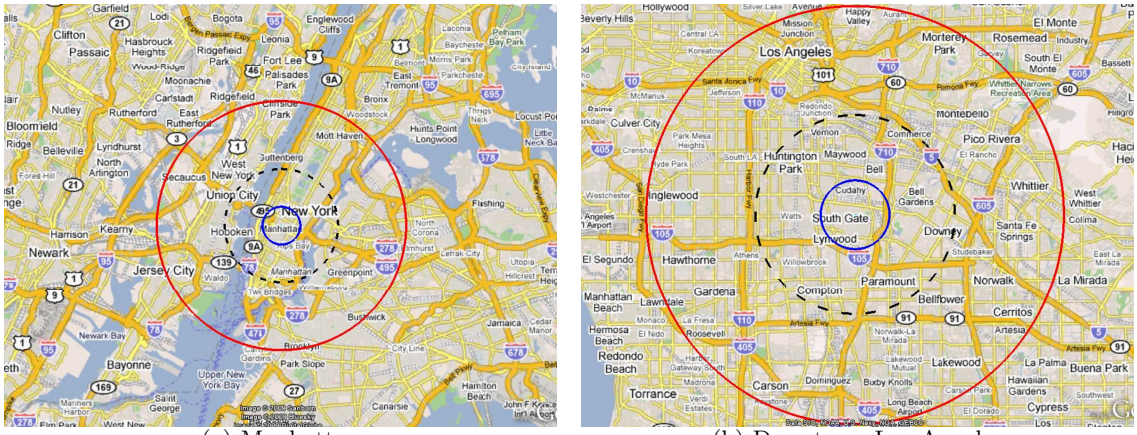
Sociologists have traditionally relied on surveys and observations of small numbers of people (e.g, airline flight paths[5]) to get a glimpse into the way that people move about. These methods often result in small sample sizes and may introduce inaccuracies due to intentional or unintentional behaviors on the part of those being observed. However, with the advent of cellular wireless communication, a ubiquitous network is now in place that must know the location of the millions of active cell phones in its coverage area in order to provide them with voice and data services. Given the almost constant physical proximity of cell phones to their owners, location data from these networks has the potential to revolutionize the study of human mobility.

In this work we explore the use of discrete location information from a cellular network to characterize human mobility. More specifically, we analyze anonymous records of approximate cell phone locations to compile aggregate statistics of how far humans travel daily in two major cities in the United States: Los Angeles (LA) and New York (NY). We introduce the concept of a *daily range*, that is, the maximal distance that a phone, and by assumption its owner, has been seen to travel in one day. We then make various observations about these daily ranges in the two populations of interest. For example, we see in Figure 1 that cell phone users in LA have median daily ranges that are nearly double those of their NY counterparts.

Our main observations from this work are as follows:

- Studying cell phone location data brings to light significant differences in mobility patterns between different human populations. One example is the large difference in median daily ranges between Los Angeles and New York residents, as mentioned above.
- Extracting a variety of statistics from this data can bring out unexpected aspects of human behavior. For example, although Angelenos' daily commutes seem to be up to two times longer than New Yorkers', New Yorkers' long-distance trips tend to be up to six times longer.
- Inspecting the data at multiple geographic granularities can further illuminate mobility patterns in different areas of the same city. For example, daily ranges across different areas of Los Angeles are more similar to each other than they are across different areas of New York.

Overall, we conclude that the study of anonymized and discrete location information from cellular networks holds great promise for the accurate and comprehensive characterization of human mobility patterns. The rest of this paper describes in more detail our data set, our data analysis methodology, and the results of our analysis.



(a) Manhattan (b) Downtown Los Angeles

Figure 1: Maps giving a visual representation of the median daily ranges of cell phone users in Manhattan and downtown Los Angeles. The radii of the inner, middle, and outer circles represent the 25^{th}, 50^{th}, and 75^{th} percentiles, respectively, of these ranges across all users in a city. Ranges for all users in a city are made to originate in a common point for clarity of display. The two maps are drawn to the same scale.

2. DATA SET

2.1 Data Characteristics

For the study described in this paper, we first developed a target set of 891 zip codes located in the New York and Los Angeles metropolitan areas. In the New York metropolitan area, these zip codes cover the five New York City boroughs (Manhattan, Brooklyn, Bronx, Queens, and Staten Island) and ten New Jersey counties that are close to New York City (Essex, Union, Morris, Hudson, Bergen, Somerset, Passaic, Middlesex, Sussex, and Warren). In the Los Angeles metropolitan area, the zip codes cover the counties of Los Angeles, Orange and Ventura. Figure 2 shows the zip codes used in the study colored in gray and black. The zip codes colored in black represent the downtown areas of the two metropolitan areas and follow the pattern of 100xx for New York City and 900xx for Los Angeles. Note that our selected zip codes cover similarly sized geographic areas in New York and Los Angeles.

We then obtained a random sample of anonymized Call Detail Records (CDRs) for 5% of the cell phone numbers where the owner's address was in one of the selected zip codes. These CDRs contained information about three types of events in the cellular network: incoming voice calls, outgoing voice calls, and data traffic exchanges. In place of a phone number, each CDR contained an anonymized identifier composed of the 5-digit zip code and a short integer. In addition, each record contained the starting time, the duration of the event, and the locations of the starting and the ending cell towers associated with the event. This random sample, generated over a 62 consecutive day period (March 15, 2009 to May 15, 2009), resulted in hundreds of thousands of anonymized identifiers, 54% from Los Angeles zip codes, and 46% from New York. The overall process yielded hundreds of millions of anonymous CDRs for analysis, for an average of 21.14 voice and data CDRs per phone per day.

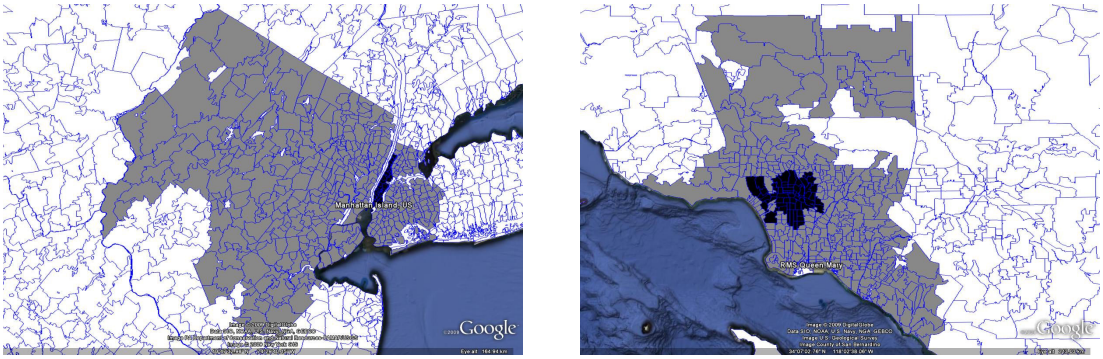
After receiving the anonymized CDRs, we checked to see if the number of identifiers in each zip code was proportional to US Census figures. In several zip codes, there were far more identifiers than expected, corresponding to accounts owned by businesses, not individuals. In order to ensure that the random sample provided an accurate representation of the populations in the zip codes, we got a list of the identifiers corresponding to businesses and omitted them from further consideration.

Since it has been found that cell phone usage amoung individuals exhibits high heterogenaity[7], we also removed from our sample records from those identifiers that rarely (less than half the days they made calls) appeared in the area of their base zip code. We assume that those users live in other parts of the country (e.g., college students), and that their commuting patterns therefore are not representative of the geographical areas we are interested in. After excluding those identifiers, we have assumed that the zip codes in our study correspond to the user's home address.

2.2 Data Anonymization and Privacy

Given the sensitivity of the data, we took several steps to ensure the privacy of individuals. First, only anonymous and aggregate data was used in this study; no information that could be used to directly or indirectly identify individual subscribers was included. Phone numbers and CDRs were anonymized to remove any identifying characteristics. Second, the results discussed in Section 4, are presented as aggregates and no individual anonymous identifier has been singled out for the study.

Third, the data in each CDR included location information only for the cellular towers with which the phone was associated at the beginning and end of a voice call or data exchange. The phones were effectively invisible to us aside from these begin and end events. In addition, we could estimate the phone locations only to the granularity of the cell tower coverage radius. These radii average about a mile, giving uncertainly of about 3 square miles for any event.



(a) Zip codes used from New York and New Jersey. Black zip codes follow the pattern 100xx.

(b) Zip codes selected form the Los Angeles area. Black zip codes follow the pattern 900xx.

Figure 2: Billing zip codes of cell phone users in this study are shown in grey and black. Black zip codes are in the downtown areas of each city, i.e., Manhattan and downtown LA.

3. METHODOLOGY

Each anonymized phone number in the CDRs has the location of the cell towers with which the phone was associated during the initialization and the termination of calls. We use these cell tower locations as an approximation of the actual phone locations.

We define a phone's *daily range* as the maximal distance the phone has traveled in a single day. We construct the phone's daily range by calculating distances between all pairs of locations visited by the phone on a given day and extracting the maximal distance. By sorting these daily ranges and extracting the median and the maximal values, for each phone, we can calculate the *median daily range* and the *maximum daily range*, respectively. Note that while the median daily range is an approximation of the "common" daily commute distance, the maximum daily range corresponds to the longest trip taken by the phone during the study.

We acknowledge that our data might not be representative of the actual commuting patterns because individuals might not necessarily make phone calls at every place they go to. However, we feel that people do tend to use their phone at places where they spend significant amounts of time.

We go further and categorize these ranges by whether they occurred on weekends or weekdays. Our reasoning is that for the majority of people, a weekday range is more closely related to business travel (e.g., commuting, business trips) while weekend travel is more often done for pleasure.

For the discussion of the results, we divided the users based on their billing zip codes. For phones registered in New York, the groups are: Manhattan, Brooklyn, Bronx, Queens, Staten Island, and New Jersey. Users in the Los Angeles area were classified as being from Downtown LA, Beverly Hills, Antelope Valley, San Fernando Valley, or Orange County.

4. RESULTS

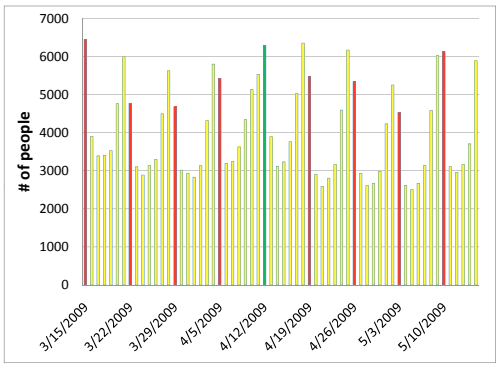


Figure 3: Number of users whose maximum daily range throughout the study falls on each date of the study. Darker bars indicate Sundays and are provided as guides for the eye only. The middle dark bar represents Easter Sunday.

Here we summarize our results with the help of boxplots, histograms, and map overlays. The boxplots in Figures 4-7 succinctly capture five-number summaries of empirical distributions of interest. The "box" represents the 25^{th}, 50^{th}, and 75^{th} percentiles, while the "whiskers" indicate the 2^{nd} and 98^{th} percentiles. The horizontal axes show miles on a logarithmic scale, and the number of people in the represented population is given in the category label.

Fridays are Weekend Days: Figure 3 is a histogram of the number of users who reach their maximum daily range on a given day of the study. These maxima occur far more frequently on Saturdays and Sundays than on weekdays with the notable exception of Fridays. When considering daily ranges, Fridays are more similar to Saturdays and Sundays and therefore we treat them as weekend days. This observation matches a similar one made by Sarafijanovic-Djukic

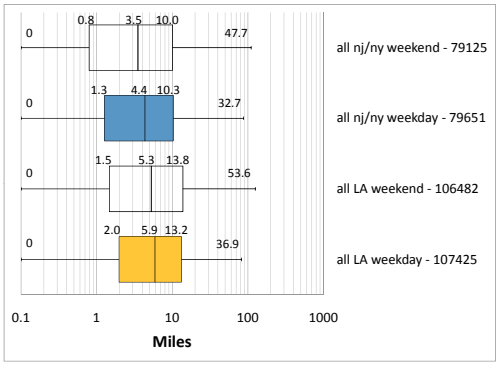


Figure 4: Boxplots of the daily ranges. Light boxes represent LA while dark boxes represent NY. Solid boxes represent weekdays and hashed boxes are the weekend.

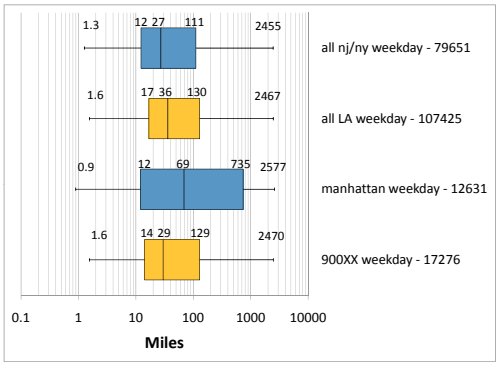


Figure 5: Boxplots of maximum daily ranges during the weekdays. Light boxes represent LA while dark boxes represent NY. The lower two boxplots represent only the city centers.

et al. [9], who eliminated weekends (including Fridays) from their own mobility study after examining the data. Further, it is of note that though data from the Easter weekend was included in the period of our study, the number of daily maxima in that weekend is not significantly larger than other weekends (although there was a slight increase in maxima in the weekdays leading up to the holiday).

Weekends are Varied: Although more daily range maxima occur on weekends, this does not necessarily correlate to greater distances traveled on weekends. As Figure 4 shows, weekends tend to be more variable than weekdays. The larger boxes corresponding to weekends can be interpreted to mean that the middle two quartiles have more variable travel patterns compared to weekdays. Specifically, the middle quartile span in miles for weekdays is $[2, 13.2]$ while for weekends it is $[1.5, 13.8]$ in LA. A possible explanation is that more people stay at home on weekends (bringing down the 25^{th} percentile) while others make longer that usual trips (bringing up the 75^{th} percentile).

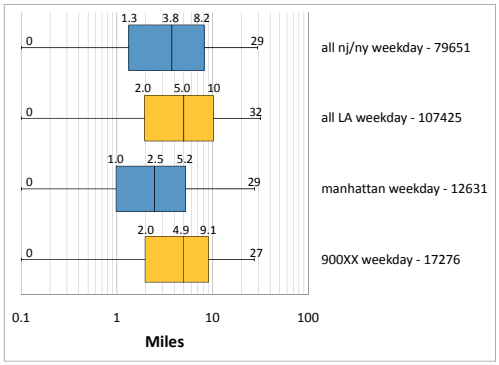
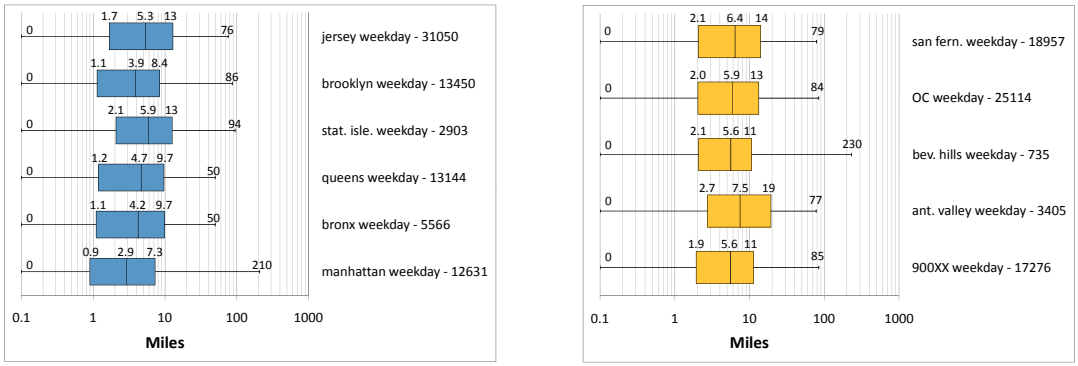


Figure 6: Boxplots of median daily ranges during the weekdays. Light boxes represent LA while dark boxes represent NY. The lower two boxplot represent only the city centers.

Angelenos Commute Farther: Comparing the travel patterns of Angelenos and New Yorkers on weekdays in Figure 4 shows non-trivial differences in human mobility. Specifically, the median for weekday daily range is 4.4 miles in NYC and 5.9 in LA, making LA daily ranges 34% larger. The 25^{th} percentile weekday numbers are 1.3 for NY and 2.0 for LA, making LA ranges 53% larger. One likely explanation for this would be that the average distance between home and work is greater in the LA area than in the NYC area. This trend of Angelenos traveling farther than New Yorkers continues when examining maximum daily ranges, as can be seen in the top two boxplots of Figure 5. The figure demonstrates that people living in the greater LA area travel about 20% farther than those from the NYC area, regardless of the percentile considered.

Commuting Estimates: We also examined the median values of users daily ranges, shown in Figure 6. Since the median daily range is the most commonly traveled distance for each user, it provides a reasonable measure of the commute-distance of users. For the greater NYC and LA regions, the medians are fairly low at 3.8 and 5.0, respectively. Using finer granularity in examining median daily range in the city centers, in the bottom two boxplots of Figure 6, confirms the general pattern of Angelenos commuting farther. Specifically, we look at detail at data from Manhattan (zipcodes 100xx) and downtown LA (zipcodes 900xx). Here we see again that Angelenos tend commute about twice father than those in New York (2x more at the 25^{th} percentile and nearly 2x at the median and at the 75^{th} percentile).

Data released by the US census [12] indicates that people living in New York City have the longest commutes in the nation by *time*. Our data supports the hypothesis that people in New York City travel significantly smaller distances than those in LA. If not necessarily contradictory, our data indicates that commuting is done differently in NYC and LA. It is possible that generally slower forms of transportation, such as public transport, or walking, might be responsible for the long commute times reported in NYC.



(a) New York area daily ranges

(b) Los Angeles area daily ranges

Figure 7: Boxplots of daily ranges, broken down into subregions of the LA and NY metropolitan areas.

City of Neighborhoods: There is further insight to be gained in breaking down the LA and NY areas, as is done in Figure 7. Even on a neighborhood-by-neighborhood level of granularity the variations are striking. Within Los Angeles, variations span from a 1.3x difference (at the median) to a difference of 3x (at the 98^{th} percentile). The differences within LA itself are, thus, equal to, or perhaps even a bit greater than, differences between LA and NYC. In New York, however, differences span from 1.8x (at the 75^{th} percentile) to 4.2x (at the 98^{th} percentile). It is easy to see that LA is more self-similar than New York with the help of the map-overlays in Figure 8.

Manhattanites Travel Very Far: Using finer granularity in examining maximum daily ranges in the city centers, in the bottom two boxplots of Figure 5, reveals an interesting reversal of general pattern of Angelenos traveling farther. Specifically, we look at detail at data from Manhattan (zipcodes 100xx) and downtown LA (zipcodes 900xx). Here the medians are at 69 and 29 for downtown NYC and downtown LA, showing that when people from New York travel far, they travel really far. At the 75^{th} percentile the numbers are 129 and 735, respectively, which is a very large difference. We note here that business phones were excluded from our dataset. However, business travel is likely to be associated with such long-distance trips as when going out of town, people are likely to take along their personal phone as well as their business phones.

USA vs Unnamed European Country: It is possible to compare some of our statistics to those computed by González *et al.* [4] for an Unnamed European Country (UEC). Our maxima show that in the greater LA area, 50% traveled more than 36 miles on at least one day and that in NY and NJ, 50% of people traveled more than 27 miles. This is in sharp contrast to González et al.'s findings that nearly 50% of all the people in their study remained within a 6 mile range over the 6-month period. The NY/NJ maxima are more than 4x larger than those in UEC and the LA maxima are more than 5x larger. While it is not surprizing that the numbers in the USA are larger, as the USA is largely car-oriented, the magnitude of the difference is unexpected.

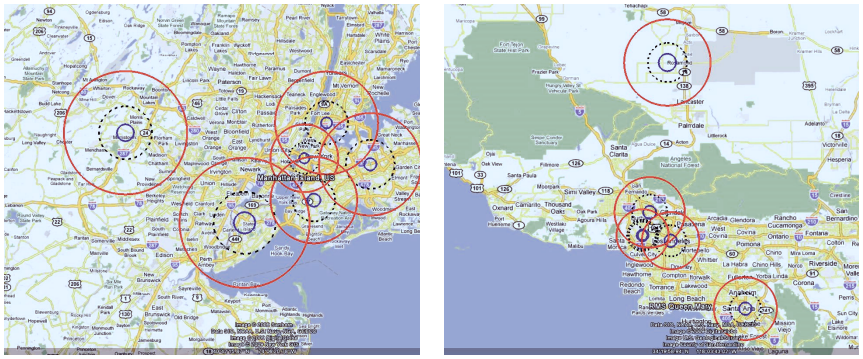
5. RELATED WORK

The usefulness of cell phone billing records has not gone unnoticed in the research community. González et al. [4] used this type of data from an unnamed European city to track people's movements for a six month period and form statistical models of the ways in which humans moved. Though the duration of the study was significantly longer than the one performed for this paper, our user base is significantly larger and we record far more user events. Further, the aims of the projects are different. González's team was interested in verifying a statistical model, while we are interested in examining whether we can observe differences in people's behavior. It is also interesting to contrast the findings of a European population with those of Americans.

Other attempts at performing studies of user mobility also tend to focus on finer grained movement patterns of individual users. Sohn et al. use GSM data to track the locations of three individuals with remarkable accuracy[10]. Similarly, Mun et al. have developed PIER[6] to track the environmental impact of individual users of the system. In contrast, our goal is to look on a more macro-scale at the ways in which cellphone users in the cities as a whole behave.

Another team, led by Girardin, used cell phones within cities to determine locations of users in New York City[3] and Rome[2]. The team found that they were able to find where people clustered and the major paths people tended to take through the cities. They were also able to find differences between the behavior of locals and tourists. In addition to cell phone records, the team relied on tagged photos uploaded to popular photo sharing websites. Unlike in our study, no effort was made to compare the movement patterns in the two cities studied. Further, we are more interested in long term aggregate behavior than the short term travel patterns studied by Girardin.

In an effort to step away from the patterns of individual users, Pulselli et al. and Ratti examine how the call volume can be used as a proxy for population density in Milan[8]. Although their work does allow for the visualization of general trends of motion through the city, we feel that we are



(a) New York and New Jersey subregions (b) Los Angeles subregions

Figure 8: Maps giving a visual representation of the median daily ranges of cell phone users in subregions of the LA and NY metropolitan areas. The radii of the inner, dashed, and outer circles represent the 25^{th}, 50^{th}, and 75^{th} percentiles, respectively, of these ranges across all users in a subregion. Ranges for all users in a subregion are made to originate in a common point for clarity of display.

able to derive a more nuanced picture of human mobility.

Even when cell phone records were not accessible, researches have come up with innovative ways to find patterns of human mobility. A team led by Tang [11] used connections to wireless access points in California and found that they were able to find subsets of users that displayed similar access patterns. These subsets were then used to analyze movement. Similarly, we examine subsets of users, but we break users down into geographic regions to find mobility patterns. We also use the cellular network which provides us much richer information regarding user locations.

6. CONCLUSIONS

Cellular phone networks can help solve important problems outside the communications domain due to their rich insights into the way people move. By analyzing aggregated records of cell phone locations, we have been able to draw novel conclusions regarding the mobility patterns of people in and around two major cities in the United States, namely Los Angeles and New York.

Using the concept of a daily range of travel, we have demonstrated concrete differences between Angelenos and New Yorkers. Those living in the area of Los Angeles tend to travel on a regular basis roughly 2 times farther than people in and around New York. However, when looking at the maximum distance traveled by each person, New Yorkers are prone to taking 2-6 times longer trips than Angelenos .Furthermore, by looking within the cities themselves, we see that although significant differences exists between portions of both cities, the LA area is more homogeneous than the New York area.

Our results to date demonstrate the potential of our approach to characterizing human mobility patterns on a large scale. In future work we plan to use such data to identify certain areas that are most frequently visited by users. Using this information we hope to more precisely quantify commute distances, and thus the impact of commuting behavior on the carbon footprints of different populations.

7. REFERENCES

[1] D. Brockmann, V. David, and A. M. Gallardo. Human mobility and spatial disease dynamics. *Proc. of the Workshop on Social Computing with Mobile Phones and Sensors: Modeling, Sensing and Sharing*, Aug. 2009.

[2] F. Girardin, F. Calabrese, F. Dal Fiorre, A. Biderman, C. Ratti, and J. Blat. Uncovering the presence and movements of tourists from user-generated content. In *Proc. of International Forum on Tourism Statistics*, 2008.

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

[4] M. C. González, C. A. Hidalgo, and A.-L. Barabási. Understanding individual human mobility patterns. *Nature*, 453, June 2008.

[5] R. Guimera and L. Amaral. Modeling the world-wide airport network. *Eur Phys J B*, 38, Jan. 2004.

[6] M. Mun, S. Reddy, K. Shilton, N. Yau, J. Burke, D. Estrin, M. Hansen, E. Howard, R. West, and P. Boda. Peir, the personal environmental impact report, as a platform for participatory sensing systems research. *Proc. of the International Conference on Mobile Systems, Applications and Services*, June 2009.

[7] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, K. Kertész, and A. Barabási. Structure and tie strengths in mobile communication networks. *Proc. of the National Academy of Sciences of the United States of America*, 104, 2007.

[8] R. Pulselli, P. Ramono, C. Ratti, and E. Tiezzi. Computing urban mobile landscapes through monitoring population density based on cellphone chatting. *Int. J. of Design and Nature and Ecodynamics*, 3, 2008.

[9] N. Sarafijanovic-Djukic, M. Piórkowski, and M. Grossglauser. Island hopping: Efficient mobility-assisted forwarding in partitioned networks. *SECON*, Sept. 2006.

[10] T. Sohn, A. Varshavsky, A. LaMarca, M. Y. Chen, T. Choudhury, I. Smith, S. Consolvo, J. Hightower, W. G. Griswold, , and E. de Lara. Mobility detection using everyday gsm traces. *Proc. of the International Conference on Ubiquitous Computing*, Sept. 2006.

[11] D. Tang and M. Baker. Analysis of a metropolitian-area wireless network. *Wireless Networks*, 8, March-May 2002.

[12] US census data. Downloaded from http://www.census.gov.