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 has great potential as a tool for studying these patterns. In this work, we use anonymous and aggregate statistics of the approximate locations of hundreds of thousands of cell phones in Los Angeles and New York City to demonstrate different mobility patterns in 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 Los Angeles counterparts.

1. INTRODUCTION

Characterizing human mobility patterns is critical to a deeper understanding of the effects of human movement. For example, the impact of human travel on the environment cannot be understood without such a characterization. Similarly, understanding and modeling the ways in which disease spreads hinges on a clear picture of the ways that humans themselves spread [2]. Other examples abound in fields like urban planning, where knowing how people come and go can help determine where to deploy infrastructure [1].

Human mobility researchers have traditionally relied on surveys and observations of relatively small numbers of people to get a glimpse into the way that humans move about, for instance by studying airline flight paths [8]. 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, ubiquitous networks are now in place that must know the location of the millions of active cell phones in their coverage areas in order to provide the phones 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.

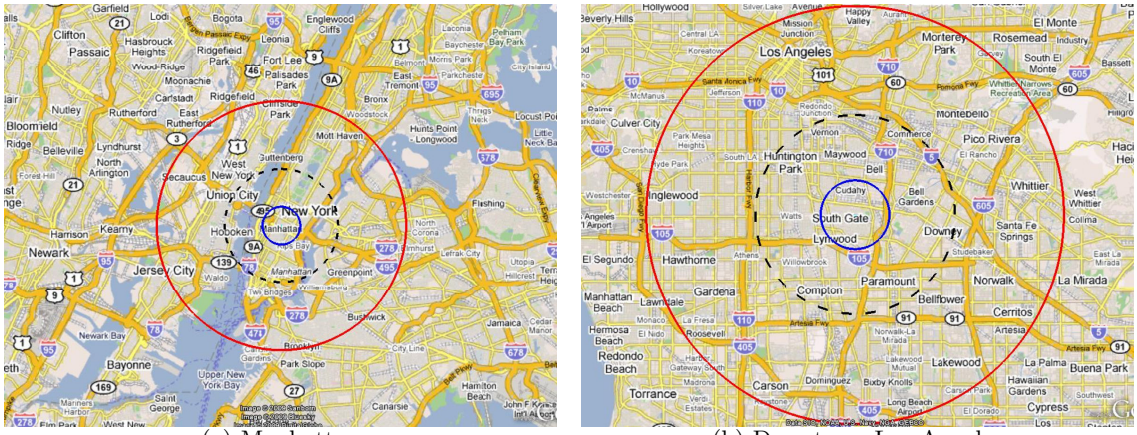
In this work, we explore the use of location information from a cellular network to characterize human mobility in two major cities in the United States: Los Angeles (LA) and New York (NY). More specifically, we analyze anonymous records of approximate cell phone locations at discrete times when the phones are in active use. Our data set spans two months of activity for hundreds of thousands of phones, yielding hundreds of millions of location events. We then compile aggregate statistics of how far humans travel daily. 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. Finally, we 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 downtown LA have median daily ranges that are nearly double those of their Manhattan 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 LA and NY 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 two times longer than New Yorkers', New Yorkers' long-distance trips seem to be 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 LA are more similar to each other than they are across different areas of NY.

Overall, we conclude that the study of operational records from cellular networks holds great promise for the large-scale characterization of human mobility patterns without compromising individual privacy. The rest of this paper describes in more detail our data set, our analysis methodology, and our results.

HotMobile 2010, February 22-23, Annapolis, Maryland, USA



(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 this study, we first developed a target set of 891 zip codes located in the New York and Los Angeles metropolitan areas. In the NY area, these zip codes cover the five New York City boroughs (Manhattan, Brooklyn, Bronx, Queens, and Staten Island) and ten New Jersey counties that are close to New York City (Essex, Union, Morris, Hudson, Bergen, Somerset, Passaic, Middlesex, Sussex, and Warren). In the LA area, the zip codes cover the counties of Los Angeles, Orange and Ventura. Figure 2 shows the zip codes used in the study. Note that our selected zip codes cover similarly sized geographic areas in NY and LA.

We then obtained a random sample of anonymized Call Detail Records (CDRs) for 5% of the cell phone numbers where the owner's billing 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 and duration of the event, and the locations of the starting and 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 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 [15] for the overall populations in these zip codes. Several zip codes had far more identifiers than expected, corresponding to accounts owned by businesses, not individuals. We omitted identifiers corresponding to businesses from further consideration.

We also removed from our sample the records for those phones that appeared in the area of their base zip code fewer than half the days they made calls. We assumed that the owners of those phones live in other parts of the country (e.g., they are college students), and therefore that their daily travel patterns are not representative of the geographical areas we are interested in. After excluding those identifiers from consideration, we have assumed that the zip codes in our records correspond to users' home addresses.

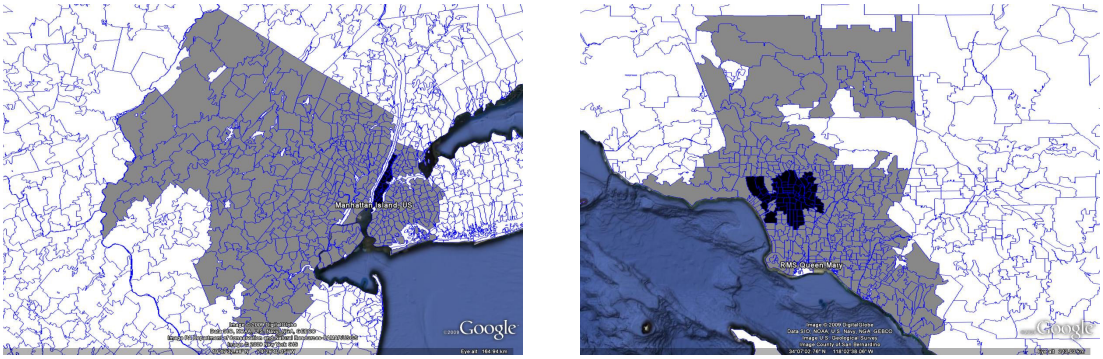
2.2 Data Anonymization and Privacy

Given the sensitivity of the data, we took several steps to ensure the privacy of individuals. First, only anonymous data was used in this study. In particular, CDRs were anonymized to remove any personally identifying characteristics. Second, all our results are presented as aggregates and no individual anonymous identifier was singled out for the study. By observing and reporting only on the aggregates, we protect the privacy of individuals.

Third, each CDR only included location information for the cellular towers with which a phone was associated at the beginning and end of a voice call or 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 an uncertainty of about 3 square miles for any event.

3. METHODOLOGY

Throughout this study, we use the locations of cellular towers with which a phone is associated as approximations of that phone's actual locations. We define a phone's *daily range* as the maximal distance it has traveled in a single day. We construct a phone's daily range by calculating distances between all pairs of locations visited by the phone on a given day, and extracting the maximal pairwise distance. Because distances are calculated "as the crow flies," our daily range is more accurately a lower bound on the maximal distance a phone has traveled. By repeating this process for each day of the study, we end up with 62 daily ranges for each phone.



(a) New York metropolitan area

(b) Los Angeles metropolitan area

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; they follow the pattern 100xx for Manhattan and 900xx for downtown LA.

By further calculating the median and maximal values of these daily ranges over the duration of our study, we arrive at a phone's *median daily range* and *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 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 work-related travel (e.g., commuting, business trips), while weekend travel is more often done for pleasure.

In addition, we divided the phones into groups based on their billing zip codes. For phones registered in the NY area, the groups are Manhattan, Brooklyn, Bronx, Queens, Staten Island, and New Jersey. Phones in the LA area were classified as being from Downtown LA, Beverly Hills, Antelope Valley, San Fernando Valley, or Orange County.

We acknowledge that our data may not represent actual commuting patterns because individuals do not necessarily use their phones in every place they go. However, we feel that people do tend to use their phones in places where they spend significant amounts of time.

4. RESULTS

We summarize our results with the help of boxplots, histograms, and map overlays. The boxplots in Figures 4-7 depict five-number summaries of the complete empirical distributions of interest. The "box" represents the 25^{th}, 50^{th}, and 75^{th} percentiles, while the "whiskers" indicate the 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.

The statistical significance of our results is apparent from our boxplots. We could have shown the variability in our data using a technique known as notched box plots[9], where the size of a notch around the median represents the variation of the median. Boxplots whose notches do not overlap would be considered to have come from distributions with significantly different medians. However, because of the large size of our data set, our notches would be imperceptibly small, about the same width as our median lines. In other words, any visible difference between our median lines is statistically significant.

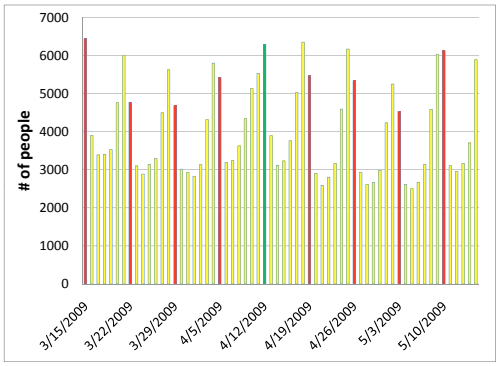


Figure 3: Number of users whose maximum daily range throughout the study falls on each date of the study. Darker bars indicate Sundays. The middle dark bar represents Easter Sunday.

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 workdays, 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 et al. [12], who eliminated weekends (including Fridays) from their own mobility study after examining their 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 max-

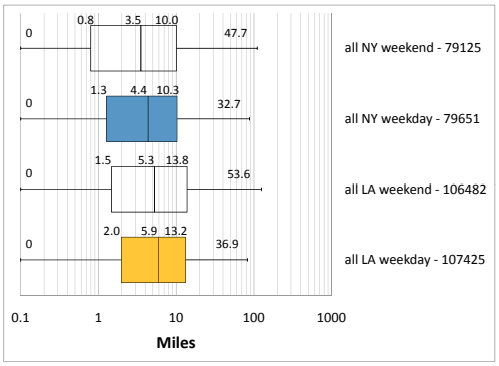


Figure 4: Boxplots of all daily ranges during weekdays and weekends. Light boxes represent LA while dark boxes represent NY.

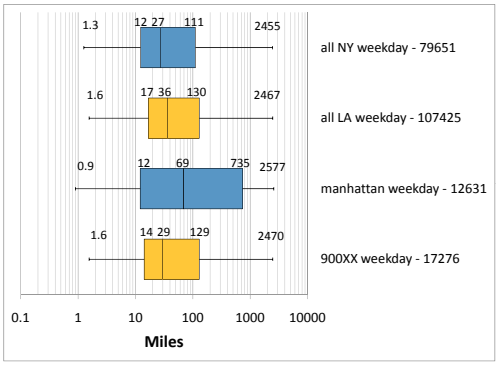


Figure 5: Boxplots of maximum daily ranges during weekdays. Light boxes represent LA while dark boxes represent NY. The lower two boxplots represent only the city centers.

ima 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, in LA the middle quartile span for weekdays is $[2, 13.2]$ miles, while for weekends it is $[1.5, 13.8]$ miles. 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).

Angelenos Commute Farther: Figure 4 shows non-trivial differences between the weekday travel patterns of Angelenos and New Yorkers. Specifically, the median for weekday daily range is 4.4 miles in NY 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 NY 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 LA area travel about 20% farther than those

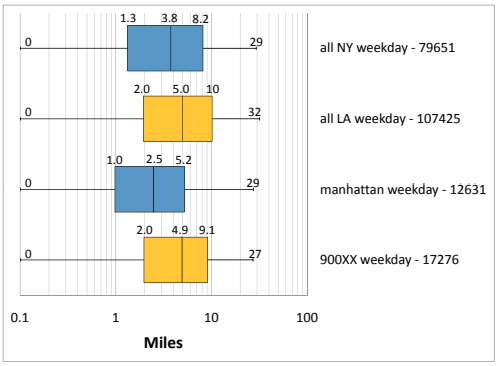


Figure 6: Boxplots of median daily ranges during weekdays. Light boxes represent LA while dark boxes represent NY. The lower two boxplots represent only the city centers.

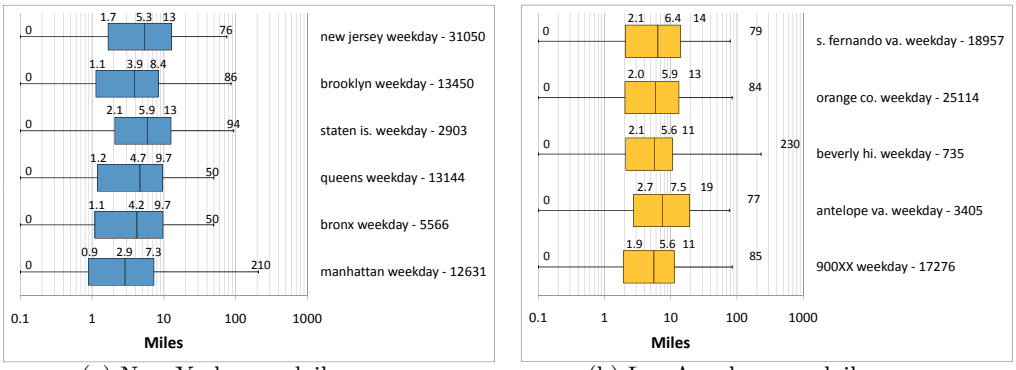
from the NY area, regardless of the percentile considered.

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 his daily commute distance. For the greater NY and LA regions, the medians are fairly low at 3.8 and 5.0 miles, respectively. Using finer granularity in examining median daily range in the city centers, the bottom two boxplots of Figure 6 confirm 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 to commute about twice as far as New Yorkers (2x at the 25^{th} percentile and nearly 2x at the 50^{th} and 75^{th} percentiles).

Data released by the US Census Bureau [15] indicates that people in NY have the longest commutes in the nation by *time*. Our data suggests that people in NY have significantly shorter commutes than people in LA by *distance*. If not necessarily contradictory, our data indicates that commuting is done differently in NY and LA. It is possible that generally slower forms of transportation, such as public transport or walking, are responsible for the long commute times reported in NY.

City of Neighborhoods: There is further insight to be gained in breaking down the LA and NY areas into subareas, as is done in Figure 7. Variations in mobility are striking even between subareas of the same city. Within LA, variations span from 1.3x (at the median) to 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 NY. Differences within NY are even greater — variations span from 1.8x (at the 75^{th} percentile) to 4.2x (at the 98^{th} percentile). The map overlays in Figure 8 also show that LA is more self-similar than NY.

Manhattanites Travel Very Far: By examining maximum daily ranges only in the city centers, the bottom two boxplots of Figure 5 reveal an interesting reversal of the general pattern of Angelenos traveling farther than New Yorkers. Specifically, we look at weekday data from Manhattan (zipcodes 100xx) and downtown LA (zipcodes 900xx). Here the medians are at 69 and 29 miles for Manhattan and downtown LA, respectively. For the 75^{th} percentiles the cor-



(a) New York area daily ranges (b) Los Angeles area daily ranges

Figure 7: Boxplots of all weekday daily ranges, split into subregions of the LA and NY metropolitan areas.

responding numbers are 735 and 129 miles. These numbers show that when Manhattanites travel far, they travel very far and farther than Angelenos. We recall that business phones were excluded from our dataset. However, business travel is still likely to be associated with these long-distance weekday trips, because when going out of town people are likely to take along their personal phones as well as their business phones.

US vs Unnamed European Country: It is possible to compare some of our statistics to those computed by González et al. for an Unnamed European Country (UEC) [7]. Our maxima show that in the greater LA area, 50% of people traveled more than 36 miles on at least one day, and that in the NY area 50% 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 a 6-month period. The LA maxima are more than 5x larger than those in UEC and the NY maxima are more than 4x larger. While it is not surprising that the numbers in the US are larger, as the US is more car-oriented, the magnitude of the difference is unexpected.

5. RELATED WORK

The usefulness of cellular network operational records has not gone unnoticed in the research community. González et al. [7] used this type of data from an unnamed European country to track people's movements for a 6-month period and form statistical models of how individuals move. Though the duration of their study was longer than ours, our user base is significantly larger and we analyze far more location events. Further, the aims of the two projects are different. González's et al. were interested in modeling an individual, while we are interested in differences in behavior between large populations. It is also interesting to contrast the mobility patterns of European and American populations, as we did in Section 4.

Other attempts at studying user mobility also tend to focus on finer-grained movement patterns of individual users. Sohn et al. used GSM data to determine mobility modes, such as walking or driving, of three individuals [13]. Similarly, Mun et al. developed PEIR [10] to track the environmental impact of individual users of the system. In contrast, our goal here has been to look on a more macro scale at the ways in which whole populations behave.

Work by Girardin et al. used cell phone usage within cities to determine locations of users in Rome [5] and New York City [6]. They were able to find where people clustered in these cities 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, they relied on tagged photos uploaded to popular photo sharing websites. In contrast to our study, they made no effort to compare movement patterns between the two cities they studied. Further, we are more interested in long-term aggregate behavior than the short-term travel patterns they studied.

In a step away from studying the patterns of individual users, Pulselli et al. examine the use of wireless call volume as a proxy for population density in Milan [11]. Although their work illuminates trends of movement through the city, we feel that our more direct measurements of user locations yield a more accurate picture of human mobility.

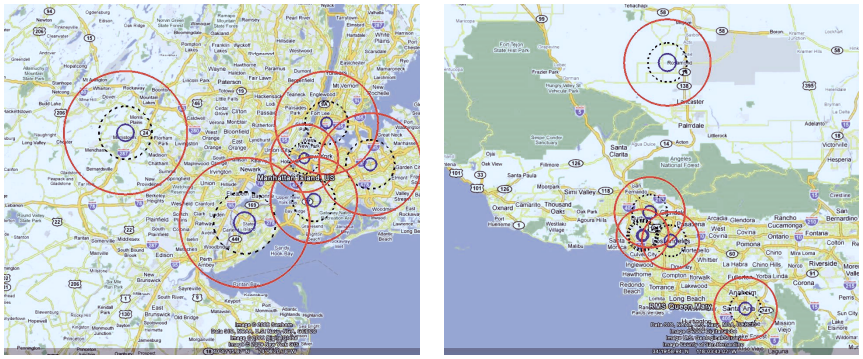
There is a growing body of work around the use of cell phones as ubiquitous sensors of factors such as location [3, 4, 10]. Our work is related to but differs from such participatory sensing efforts. One, our data comes from the cellular network, not from sensors on the phones. Two, our dataset is much larger than those used by such efforts to date.

Before cell phones became ubiquitous, Tang et al. [14] studied user accesses to an early wireless data network in California. They found subsets of users with similar access patterns and analyzed movement within these subsets. Our study is based on much larger numbers of people in multiple geographic regions, which makes our results more representative of the population at large.

6. CONCLUSIONS & FUTURE WORK

Cellular phone networks can help solve important problems outside the communications domain because they can provide rich insights into the way people move. Scientists and policy makers in many fields can use human mobility data to explore existing problems and anticipate future problems. By analyzing anonymized records of cell phone locations, we have been able to draw novel conclusions regarding how people move in and around two major cities in the United States, Los Angeles and New York.

Using the concept of a daily range of travel, we have demonstrated concrete differences between Angelenos and



(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.

New Yorkers. Those living in the LA area tend to travel on a regular basis roughly 2 times farther than people in and around NY. 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 NY area.

Our results to date demonstrate the potential of our approach to characterizing human mobility patterns on a large scale and without compromising individual privacy. Our methodology has wide-ranging applications. Future work includes using these techniques to examine correlations between human movements and world events such as national holidays and disease outbreaks. A better understanding of how such events affect human movement can inform a range of pursuits, from urban planning to disaster response. Further, we plan to expand upon our comparison of two urban populations to a larger set of population types, for example to compare rural vs. urban movement patterns. Finally, we plan to use clustering techniques to identify those areas most frequently visited by cell phone users. We thus hope to more precisely quantify commute distances, and thereby the impact of commuting behavior on the carbon footprints of different populations.

Acknowledgments

We thank our shepherd Ben Greenstein and our anonymous referees for comments that helped to improve this paper.

7. REFERENCES

[1] The journey to work: Relation between employment and residence. Technical Report No. 26, American Society of Planning Officials, May 1951.

[2] 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.

[3] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava. Participatory sensing. In *Workshop on World-Sensor-Web (WSW06): Mobile Device Centric Sensor Networks and Applications*, pages 117–134, 2006.

[4] D. Cuff, M. Hansen, and J. Kang. Urban sensing: out of the woods. *Commun. ACM*, 51(3):24–33, 2008.

[5] 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.

[6] 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.

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

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

[9] R. McGill, J. W. Tukey, and W. A. Larson. Variations of box plots. *The American Statistician*, 32, Feb. 1978.

[10] 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.

[11] 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.

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

[13] 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.

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

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