Putting Big Data in Its Place: Understanding Cities and Human Mobility with New Data Sources

by

Jameson Lawrence Toole

Submitted to the Engineering Systems Division
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2015

Author .....
Engineering Systems Division
May 15, 2015

Certified by .....
Marta C. Gonzàlez
Associate Professor of Civil and Environmental Engineering
Thesis Supervisor

Certified by .....
Joseph M. Sussman
JR East Professor of Civil and Environmental Engineering and
Engineering Systems
Thesis Supervisor

Certified by .....
P. Christopher Zegras
Associate Professor of Transportation and Urban Planning
Thesis Supervisor

Accepted by .....
Munther Dahleh
William A. Coolidge Professor of Electrical Engineering and Computer Science
Acting Director, Engineering Systems Division

Putting Big Data in Its Place: Understanding Cities and Human Mobility with New Data Sources

by

Jameson Lawrence Toole

Submitted to the Engineering Systems Division
on May 15, 2015, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

According the United Nations Population Fund (UNFPA), 2008 marked the first year in which the majority of the planet's population lived in cities. Urbanization, already over 80% in many western regions, is increasing rapidly as migration into cities continue. The density of cities provides residents access to places, people, and goods, but also gives rise to problems related to health, congestion, and safety. In parallel to rapid urbanization, ubiquitous mobile computing, namely the pervasive use of cellular phones, has generated a wealth of data that can be analyzed to understand and improve urban systems. These devices and the applications that run on them passively record social, mobility, and a variety of other behaviors of their users with extremely high spatial and temporal resolution. This thesis presents a variety of novel methods and analyses to leverage the data generated from these devices to understand human behavior within cities. It details new ways to measure and quantify human behaviors related to mobility, social influence, and economic outcomes.

Thesis Supervisor: Marta C. Gonzàlez
Title: Associate Professor of Civil and Environmental Engineering

Thesis Supervisor: Joseph M. Sussman
Title: JR East Professor of Civil and Environmental Engineering and Engineering Systems

Thesis Supervisor: P. Christopher Zegras
Title: Associate Professor of Transportation and Urban Planning

Acknowledgments

In his acceptance speech for a teaching award from the Mathematics Associate of America, Harvey Mudd College Professor Francis Edward Su described living and teaching with grace. Living and teaching with grace, he explained, means understanding that your accomplishments are not what make you a worthy human being. We only realize this when we have received grace, or been given good things that we did not deserve.

Though it seems obvious to state, academic pursuits are often difficult and intimidating. I certainly did not feel worthy standing in front of a professor's door asking for a summer research position with no previous experience on my resume. Nor did I feel worthy telling a professor that I wanted to pursue research in his field after receiving an A- in his class. Luckily, these teachers and many others showed me grace. In high school, two physics teachers taught me to love the elegant, intuitive models of physics and inspired me to pursue the science further. As a sophomore undergraduate, I can still remember the watershed lecture in my differential equations class when he digressed into a tangent about modeling the economy as a system of coupled ODEs and I realized that it was possible to combine multiple disciplines I loved. Whether it was Robert Deegan, who took a risk inviting a sophomore physics major with no experience to work in his research lab for a summer, or Scott Page who invited me to tag along to a political science workshop as the only undergraduate in the room, or Nathan Eagle who mentored me during the summer I worked as a UROP at the Sante Fe Institute, I have been extraordinarily fortunate to have benefited from teachers and mentors who welcomed me into their offices and laboratories despite my sometimes embarrassing lack of credentials.

These gifts of grace were never were never more needed or appreciated throughout my graduate studies. To my advisor, Marta Gonzalez, I cannot tell you how thankful I am for the freedom you offered for me to pursue my passion and for the help you provided when I found myself stuck. To my committee, I thank you for your willingness to branch out to new areas with me and for keeping my work grounded

with your expertise. I am certain that I would not be as proud of this work without your patience and guidance.

Above all, I would like to thank my family and friends for their unconditional and unwavering support. Your encouragement in the valleys of these past five years and your celebrations at the peaks gave this journey meaning and for that I am sincerely grateful.

Contents

1 Introduction - The Science of Cities **29**

2 Human Mobility **35**

2.1 New Data Sources 36

2.2 Individual Mobility Models 40

2.3 Aggregate Mobility 47

2.4 Human Mobility and City Science 53

2.4.1 Social Behavior and Choice 53

2.4.2 Economics Behavior in Cities 54

2.5 Mobility and City Infrastructure 56

2.6 Acknowledgements 57

3 Social Behavior and Choice **59**

3.1 Introduction 59

3.2 Materials and Methods 61

3.2.1 Data 61

3.2.2 Social and Mobility Measurements 63

3.3 Results 65

3.3.1 Correlations between social behavior and mobility 65

3.3.2 Contextualizing social contacts with mobility 68

3.3.3 Coupling social ties and mobility 69

3.4 Discussion 73

3.5 Acknowledgments 74

3.6 Chapter 3 - Appendix . . . 75
3.6.1 Data . . . 75
3.6.2 Social Network Extraction . . . 75
3.6.3 Metric Definitions . . . 75
3.6.4 Controlling for number of calls . . . 78
3.6.5 Controlling for Degree . . . 79
3.6.6 Social Distance and Geographic Similarity . . . 79
3.6.7 Clustering . . . 79
3.6.8 Ego-network link type mixes . . . 80
3.6.9 Model Comparisons . . . 88

4 Economic Behavior in Cities **93**
4.1 Measuring the Economy . . . 95
4.2 Predicting the Present . . . 96
4.3 Data . . . 97
4.4 Observing Unemployment at the Community Level . . . 98
4.5 Observing Unemployment at the Individual Level . . . 99
4.5.1 Validating the Layoff . . . 100
4.6 Assessing the Effect of Unemployment at the Individual Level . . . 101
4.7 Observing Unemployment at the Province Level . . . 103
4.8 Discussion . . . 107
4.9 Acknowledgments . . . 109
4.10 Chapter4 - Appendix . . . 109
4.10.1 Materials and Methods . . . 109
4.10.2 Filtering CDR Data . . . 110
4.10.3 Manufacturing plant closure . . . 110
4.10.4 Town Level Structural Break Model . . . 111
4.10.5 Individual Structural Break Model . . . 112
4.10.6 Bayesian Estimation . . . 113
4.10.7 The European Labor Force Survey . . . 113

4.10.8 The Effect of Job Loss on Call Volumes . . . 114
4.10.9 Measuring Changes . . . 115
4.10.10 Predicting Province Level Unemployment . . . 116
4.10.11 Mass Layoffs and General Unemployment . . . 120

5 The Path Most Traveled : Estimating Travel Demand With Big Data Resources 131

5.1 Introduction . . . 131
5.1.1 Description of Data . . . 133
5.2 System Architecture and Implementation . . . 136
5.2.1 Architecture . . . 136
5.2.2 Parsing, Standardizing, and Filtering User Data . . . 136
5.2.3 Creating and storing geographic data . . . 138
5.3 Estimating Origin-Destination Matrices . . . 139
5.3.1 Measuring Flow . . . 140
5.3.2 Trip Assignment . . . 144
5.4 Results . . . 148
5.4.1 Trip Tables and Survey Comparison . . . 148
5.4.2 Road Network Analysis . . . 150
5.4.3 Bipartite Road Usage Graph . . . 150
5.4.4 Visualization . . . 155
5.5 Limitations and Future Work . . . 156
5.6 Conclusions . . . 157
5.7 Acknowledgments . . . 157
5.8 Chapter 5 - Appendix . . . 158
5.8.1 Algorithms . . . 158

6 Inferring land use from mobile phone activity and points of interest 163

6.1 Introduction . . . 163
6.2 Mobile phones and human mobility . . . 166
6.3 Data sources . . . 168

6.4 Common spatial representation . . . 169
6.5 Descriptive Statistics . . . 170
6.6 Classifying Land Use by Mobile Phone Activity . . . 174
6.7 Incorporating Points of Interest . . . 180
6.8 Conclusion . . . 182
6.9 Acknowledgements . . . 184

7 Failures in markets for personal data assets. 187
7.1 Introduction . . . 187
7.2 Market Failures . . . 190
7.2.1 Uncertain Property Rights . . . 190
7.2.2 Information deficiencies and Asymmetries . . . 195
7.2.3 Externalities . . . 198
7.3 Private solutions . . . 202
7.4 Government Regulation . . . 206
7.5 Conclusions . . . 211

8 Conclusion 215
8.1 Social Behavior and Choice . . . 216
8.2 Economic Behavior in Cities . . . 218
8.3 Mobility and City Infrastructure . . . 219
8.4 Future Directions . . . 221

List of Figures

2-1 Mobile phones are increasingly being used to collect high-resolution mobility data. This figure from de Montjoye et al. [71] depicts A) a sequence of calling events made by a user at different locations. B) These events are localized to the area served by the closest mobile phone tower to the use and C) can be aggregated into individual specific neighborhoods where a user is likely to be found at different times of the day or week. 37

2-2 A) Individual mobility trajectories are passively collected from mobile devices [97]. B) Measuring the distribution of radius of gyrations, r_g within a population of 100,000 users in a European country reveals considerable heterogeneity in typical travel distance of individuals. Moreover, this distribution cannot be explained by modeling each individual's movement as realizations of a single Levy flight process [97]. C and D) Show the slower than linear growth in new locations visited over time $S(t)$ and that the probability a location is visited next is inversely proportional to the frequency it has been visited in the past [202]. E) This preferential return contributes to strikingly high predictability $R(t)$ over time while F) the number of unique locations visited in any given hour is highly periodic and corresponds to the sleep-wake cycles of individuals [205]. 40

2-3 A) Removing geographic coordinates from locations and only focusing on a set of unique places and the directed travel between them, mobility motifs reveal that the daily routines of people are remarkably similar. Despite over 1 million unique ways to travel between 6 or fewer points, just 17 motifs are used by 90% of the population. Moreover, the frequency of their appearance in CDR data matches very closely with more traditional survey methods [190]. B) Despite this similarity and predictability, our movement displays a high degree of unicity. Just four spatiotemporal points is enough to differentiate a user from 95% of all others individuals [71]. 45

2-4 A) The radiation model accounts for intervening opportunities, producing more accurate estimates of flows between two places than more traditional gravity models [200]. B) Routing millions of trips measured from CDR data to real road networks makes it possible to measure the importance of a road based on how many different locations contribute traffic to it, K_{road}. Understanding how transportation systems perform under different loads presents new opportunities to solve problems related to congestion and make infrastructure more efficient [229]. . . . 48

2-5 A) Global air travel has dramatically increased the speed at which diseases can spread from city to city and continent to continent [163]. B) Mobility also impacts social behavior as we are far more likely to be friends with someone who lives nearby than far away [136]. C) Mobility and the access it provides has strong correlations with economic outcomes. Children have dramatically different chances at upward economic mobility in certain places of the United States than others [55]. 52

3-1 A small sample of calls between residents is shown for each of three cities. CDRs provide the location of each caller as well as record of communication between then. A dot is drawn at the approximate location of a user and a link appears between two users calling each other. Our aim is to identify useful and reproducible patterns from this coupled tangle of social and spatial behavior. 62

3-2 Similarity of visitation patterns between nodes in social networks. For each user, we keep track of (A) how many visits are made to locations across the city and (B) construct a social network by tracking calls to others. We can then define (C) the geographic cosine similarity between two users by computing the cosine of the angle between any two vectors in the location space. 62

3-3 Correlations between mobility and social behavior. For each city, we compute the (A) distribution of cosine similarity and (B) predictability using observed edges (colored lines) and compare to distributions made using randomized edges. We find both mobility similarity and predictability are much higher when using actual social contacts compared to random users. Social similarity is also correlated with mobility similarity. (C) Ranking each user's contacts by number of calls, we find that stronger ties are more geographically similar. (D) Moreover, the more common contacts shared by two users, the more geographically similar those individuals tend to be. Finally, we explore how social behavior is correlated with mobility. (E) We find that users with more unique contacts tend to visit more unique locations. (F) Users who distribute their calls to contacts more evenly (higher entropy) are more predictable than users with more uneven call distributions. This suggests that users who share social attention more evenly also share locations. Figure S2 and S3 in the Appendix 3.6 show these results controlling for call frequency. 66

3-4 Characterizing social ties based on similarity of movement over time. (A) We perform k-means clustering on the set of similarity time series from edges in the network. We find three groups emerge in each city: (*i*) *acquaintances* who have low levels of similarity across all times, (*ii*) *co-workers* who have elevated similarity during work hours on weekdays, but lower levels on weekends, and (*iii*) *family/friends* who have high similarity on nights and weekends. (B) For each city we construct subgraphs containing only edges in a single cluster. We find that these subgraphs retain high clustering coefficient (C_g) within the co-worker and family/friend group while acquaintances are far less likely to have ties among each other. Finally, we explore how an user's behavior correlates with the mobility characteristics of their immediate social network. (C-D) We group nodes based on their mobility characteristics (unique locations visited S and predictability $\frac{|\hat{\mathbf{v}}|}{|\mathbf{v}|}$) then compute the fraction of edges that belong to each of the identified clusters for each node in the group. Individuals that are more exploratory (visit more unique places) tend to have higher fractions of *acquaintances* ties than individuals with lower mobility while the reverse trend is observed for the most predictable individuals. 70

3-5 A schematic description of the GeoSim model. As in the IM model presented by Song et al., individuals first decide whether to return to a previously visited location or explore a new location. The actual choice of location to visit, new or returning, is made based on either a social influence with probability α or individual preference with probability $1-\alpha$. 81

3-6 Comparing social mobility models. A) We compare model results simulating the rate of exploration $S(t)$ compared to empirical data. While all three models appear to estimate more absolute locations visited, the rate of this growth is consistent between them and in-line with data. B) For each user, we sort locations based on the number visits and compute the frequency that a user visits a location of rank k. We find that the IM models and our extension to it reproduce this distribution well, while the TF model is much flatter, distributing visits more evenly over all locations. C) Only the GeoSim model is able to reproduce patterns of mobility similarity and D) predictability. The TF model results shown in the inset in C shows similarity values orders of magnitude below the observed data. As the similarity is heavily influenced by the frequency distribution of visits, this deviation is likely due to the flatter distribution of f_k produced by the TF model. . . . 82

3-7 Distributions of different variables (columns) for each of the three regions (rows) for groups of users with different numbers of total calls. To ensure that measurements are not simply artifacts of differences in the amount a users interacts with their phone, we plot distributions of variables for groups of users with different activity levels. Users are binned first according to the number of records they have in the data set, then distributions of various mobility and social metrics are plotted for each user group. In general calling frequency does not affect these distributions with the exception of the number of unique locations visited where the mean is shifted right for users with more calls. 83

3-8 Various correlations metrics related to social behavior and mobility while controlling for the number of calls made by each user. Again, we bin users based on the number of records they have in our data set and then measure correlations between social and mobility metrics. We find, as was the case with distributions, these correlations are unaffected by sampling frequency. 83

3-9 Correlation between the entropy of a node's call frequency distribution to contacts and mobility variables may be affected by the degree of each node. Measures such as entropy and predictability will naturally be affected by the number of contacts each user has. For example if a user contacts for people, the maximum entropy of the distribution of call frequencies to those individuals will naturally be higher than a user who has few friends. To ensure our correlations are not artifacts of the number of contacts each user has, we plot these correlations for groups of users with the same degree and show that these relations still hold. 84

3-10 Social distance and geographic similarity. Nodes who contact each other are far more similar to each other than two randomly selected nodes. Here we compute the average mobility similarity between nodes separated by a certain number of hops. Even for nodes separated more two or three hopes, we elevated levels of similarity when compared to two randomly selected nodes in the network. 85

3-11 The silhouette score for different numbers of clusters. The silhouette score is a measure of the ratio between intra- and inter-cluster variance that gives a rough measure of the quality of clustering results (higher is better). The score drops steadily from the chosen number of clusters, 3, indicating that little is gained by additional splitting. 85

3-12 k-Means clustering results for various values of k in city R1. We perform k-means clustering for multiple values of k as a manual check that our choice of 3 clusters is appropriate. In general, additional clusters appear to be variations of three main themes used in the main text. . 86

3-13 Results from a hierarchical, agglomerative clustering algorithm with Ward linkage. This clustering method clusters nodes based on connecting data points together if they are within some distance of one another and then examining connected components. The clusters in each region closely match results from k-means, suggesting that our results are robust to the exact clustering algorithm used. 87

3-14 The mix of a user's ego network versus the number of calls they make. To ensure that the relationship between the mix of a user's ego network and their mobility isn't simply due to the fact that user's with different numbers of records having different edge mixes, we plot the average make-up for suers with different numbers of calls. We find that regardless of a user's calling frequency, the makeup of their social contacts is stable. 88

3-15 Our extended mobility model. (A) A diagram showing the choices made by an individual in deciding where to move next. We compare simulation results for different values of social influence α to distributions of (B) similarity and (C) predictability found in real data. The bimodal similarity distribution is recovered for higher values of α while the predictability results suggest that this parameters may vary for individual to individual resulting in a mix among the whole population. 91

4-1 A schematic view of the relationship between job loss and call dynamics investigated at both micro and macro level. 95

4-2 Identifying the layoff date. A) Total aggregate call volume (black line) from users who make regular calls from towers near the plant is plotted against our model (blue). The model predicts a sudden drop in aggregate call volume and correctly identifies the date of the plant closure as the one reported in newspapers and court records. B) Each of the top 300 users likely to have been laid off is represented by a row where we fill in a day as colored if a call was made near the plant on that day. White space marks the absence of calls. Rows are sorted by the assigned probability of that user being laid off according to our Bayesian model. Users with high probabilities cease making calls near the plant directly following the layoff. C) Plotting the fraction of calls made near the plant by users assigned to the top decile in probability of being unemployed (red) we see a large against a control group of individuals believed to be unaffected (blue), we see as sharp, sustained drop following the closure date. Moreover, we see that laid off individuals have an additional drop off for a two week period roughly 125 days prior the plant closure. This time period was confirmed to be a coordinated vacation for workers providing further evidence we are correctly identifying laid off workers. 99

4-3 Changes in social networks and mobility following layoffs. We quantify the effect of mass layoffs relative to two control groups: users making regular calls from the town who were not identified as laid off and a random sample of users from the rest of the country. We report monthly point estimates for six social and three mobility behaviors: A) Total calls, B) number of incoming calls, C) number of outgoing calls, D) Fraction of calls to individuals in the town at the time of the call, E) number of unique contacts, and the fraction of individuals called in the previous month who were not called in the current month (churn), G) Number of unique towers visited, H) radius of gyration, I) average distance from most visited tower. Pooling months pre- and post-layoff yield statistically significant changes in monthly social network and mobility metrics following a mass layoff. J) Reports regression coefficient for each of our 9 dependent variables along with the 66% and 95% confidence intervals. 102

4-4 Predicting unemployment rates using mobile phone data. We demonstrate that aggregating measurements of mobile phone behaviors associated with unemployment at the individual level also predicts unemployment rates at the province level. To make our forecasts, we train various models on data from half of the provinces and use these coefficients to predict the other half. Panel A compares predictions of present unemployment rates to observed rates and Panel B shows predictions of unemployment one quarter ahead using an AR1 model that includes co-variates of behaviors measured using mobile phones. Both predictions correlate strongly with actual values while changes in rates are more difficult to predict. The insets show the percent improvement to the RMSE of predictions when mobile phone co-variates are added to various baseline model specifications. In general, the inclusion of mobile phone data reduces forecast errors by 5% to 20%. 104

4-5 We plot the distribution of break dates for the structural break model estimated for individuals. We find a strong, statistically significant peak centered on the reported closure date (red) with far fewer breaks on other, placebo dates. This is consistent with both our community wide model as well as the Bayesian model presented above. 112

4-6 Identifying affected individuals. A) Each user is represented by a row where we fill in a day as colored if a call was made near the plant on that day. White space marks the absence of calls. Rows are sorted by the assigned probability of that user being laid off. B) A closer view of the users identified as mostly to have been laid off reveals a sharp cut off in days on which calls were made from the plant. C) An inverse cumulative distribution of assigned probability weights. The insert shows an enlarged view at the probability distribution for the 150 individuals deemed most likely to have been laid off. 127

4-7 A timeline showing the various data collection and reporting periods. Traditional survey method perform surveys over the course of a single week per quarter, asking participants about their employment status during a single reference week. Unofficial survey results, subject to revision are then released a few weeks following the end of the quarter. Mobile phone data, however, is continually collected throughout the quarter and is available for analysis at any time during the period. Analysis of a given quarter can be performed and made available immediately following the end of the month. 128

4-8 Predicting unemployment rates using mobile phone data from only the first 6 weeks of each quarter. We follows the same procedure as the main text. Panel A compares predictions of present unemployment rates to observed rates and Panel B shows predictions of unemployment one quarter ahead using a simple AR1 model that includes co-variates of behaviors measured using mobile phones. Both predictions correlate strongly with actual values while changes in rates are more difficult to predict. The insets show the percent improvement to the RMSE of predictions when mobile phone co-variates are added to each of four traditional forecasting models. In general, mobile phone data reduces forecast errors by 3% to 6%. 129

4-9 The average values of %RSD against the number of samples per province for different features. For all features, the %RSD's decrease rapidly with sample size and stabilize to relative small values before $k = 2000$. 130

4-10 Correlations between mass layoff events and general unemployment. Using BLS data, we plot various correlations between the number of mass layoff events, the number of initial unemployment claimants due to these events, general unemployment claims, and the unemployment rate. We find strong correlation between all of these variables suggesting that mass layoffs are a good proxy for general unemployment shocks, at least at the predictive level. 130

5-1 A flowchart of the system architecture. 136

5-2 Our efficient implementation of the incremental traffic assignment (ITA) model. A sample OD matrix is divided into two increments and then split into two independent batches each. 146

5-3 Correlations between OD matrices produced by our system and those derived from travel surveys at the largest spatial aggregation of the two models. In Boston, this is town-to-town, in San Francisco, MTC superdistrict-to-super district, in Rio, census superdistrict-to-superdistrict, and in Lisbon, freguesia-to-freguesia. The larger of these area units (e.g. towns in Boston), the better our correlations, while correlations at the smallest aggregates(e.g. freguesias in Portugal), correlations are lower. However, more work must be done to understand uncertainties in estimates provided by both models. 150

5-4 Distributions of travel volume assigned to a road and the volume-over-capacity (V/C) ratio for the five cities. The values presented in the legend refers to the fraction of road segments with $V/C > 1$. 151

5-5 A graphical representation of the bipartite network of roads and sources (census tracts), with edge sizes mapping the number of users using the connected road in their individual routes. 152

5-6 Maps depicting the proposed road classification, summarized in the legend, for the five subject cities. 153

5-7 Distributions of k_{road} and k_{source} for the five cities. Inset: The unitized collapsed normal distribution for k_{source}. 154

5-8 Two screen images from the visualization platform. (a) The trip producing (red) and trip attracting (blue) census tracts using Cambridge St., crossing the Charles River in Boston. (b) Roads used by trips generated at the census tract including MIT. 154

6-1 Zoning regulation for the Boston area. Color code: orange - Residential, red - Commercial, gray - Industrial, blue - Parks, green - Other. 164

6-2 To improve computational efficiency and reconcile all mobile phone and traditional data sources, we create a uniform grid over the city. Zoning polygons (right), are rasterized to cells 200m by 200m in size (left). For cells where more than one zoning class exists, the most prevalent class is used. Given the small size of these cells, this data transformation provides an accurate map of the city while improving computational efficiency. 171

6-3 (a) Plots are shown for three different time series of average mobile phone activity within each of five land use. The first plot shows absolute activity (number of calls and SMS messages). The second plot displays z-scored time series. The bottom plot shows residual activity.(b) More detailed view of average (over cells of the same zoning class) residual activity. 173

6-4 Spatial distribution of absolute and residual phone activity over the course of a day. While absolute mobile phone activity is dominated by population density and sleep and wake patterns, residual activity reveals flows into and out of the city center over the course of a day. . 174

6-5 The left panel displays the region-wide frequency of the 20 most common point of interest tags. While some descriptors such as 'establishment' are vague, others like 'park' align closely with official zoning classifications. The right panel shows the spatial density of all POIs across the region. The central business district predictably shows the highest density, while other clusters are dispersed throughout the region.175

6-6 (a) Shows the inputs to each decision tree $h(\mathbf{x}, \theta_k)$. A time series of residual phone activity, $\mathbf{x}$, is input and activity at a random subset of times , θ_k (denoted by the blue bars), is chosen to make comparisons. (b) A depiction of the random forest shows a number of different trees making predictions based on a different set of random times. Each tree casts a weighted vote for a certain classification. A final classification, $\hat{c}$, is made by counting these votes. 176

6-7 Left plot: zoning map as predicted from mobile phone data using the random forest classification algorithm. Right plot: spatial distribution of where the algorithm predicts land use correctly and where it fails. In general, these errors seem randomly distributed in space, suggesting that errors are not the result of some spatial correlations such as population density. For comparison to actual zoning, see the left panel of Figure 6-2. 178

6-8 The left plot shows the city zoning map with residential areas removed as predicted from mobile phone data using the random forest classification algorithm. The right map displays the spatial distribution of where the algorithm predicts land use correctly and where it fails. Without residential areas to predict, the algorithm performs significantly better at predicting other uses. For comparison to actual zoning, see the left panel of Figure 6-2. 180

6-9 An analysis of classification errors. We consider three groups: (I) Cells correctly predicted to be a given use (II) Cells of a given use incorrectly predicted to be some other use (III) Cells of some other use incorrectly predicted to be a given use. For example, Group I includes all residential areas correctly predicted to be residential. Group II, residential cells predicted to be some other use (i.e. Commercial), have average activity that is the inverse of Group I, suggesting these locations were misclassified because they display fundamentally different activity patterns. Group III represent cells of other uses such as Commercial that behave like Residential. 185

List of Tables

3.1 Basic statistics on the networks and spatial extent of each region considered. 75

4.1 Regression Results - Social and Mobility Measures. 121

4.2 Correlation coefficients between normalized, aggregated calling behaviors and unemployment rates the province level. 122

4.3 PCA results for call variables. 122

4.4 PCA Loadings. Significant elements are bolded. 122

4.5 Predicting Present Unemployment Rates - Cross Validation Model Coefficients 123

4.6 Predicting Future Unemployment Rates - Cross Validation Model Coefficients 124

4.7 RMSE with the addition of CDR data from the entire quarter with various fixed effects. The best performing model is bolded. 124

4.8 RMSE with the addition of CDR data from the half of the quarter with various fixed effects. The best performing model is bolded. . . . 125

4.9 Correlations between general unemployment and unemployment resulting from mass layoffs. 125

4.10 Correlations between mass layoff and unemployment and the state level.126

5.1 A comparison of the extent of the data involved in the analysis of the subject cities. 135

5.2 Trip tables estimates. Where possible, our results are compared to estimates made using travel surveys. For each city, we report the number of person trips in millions for a given purpose or time. Trip purposes include: home-based word (HBW), home-based other (HBO), and non-home-based (NHB). Trip periods include: 7am-10am (AM), 10am-4pm(MD), 4pm-7pm (PM), and the rest of the day (RD). We note that the exact boundaries of the surveys do not exactly coincide with those used in our estimation so direct comparisons are not exact. In general, trip magnitudes align closely, with the exception of Rio de Janeiro, where the survey results report far too few trips, illustrating the difficulty of obtaining sensible measurements via certain techniques. No comparisons could be found for Porto. 149

6.1 Tabulation of Boston zoning. The land use profile of the city is dominated by residential use accounting for nearly 75%. Other uses share roughly the same percentage of remaining land. 170

6.2 Random forest classification results. The threshold refers the total number of phone events required in each cell over period of data collected to be considered for classification. Total accuracy is defined as the fraction of correctly classified cells. The share refers to the percentage of cells actually zoned for each class of use. Element (i, j) of the confusion can be interpreted as the fraction of actual zoned uses of class i that were classified as use j by the random forest. Thus the high percentages in the Res column can be interpreted as the algorithm heavily favoring classification as residential due to its overwhelming share of overall uses. 179

6.3 Random forest classification results - less residential. In this case, residential land has been removed from consideration. The algorithm is now able to correctly predict much larger fractions of rarer land uses. 181

6.4 Random forest classification results - POI Only. 182

6.5 Random forest classification results - POI and CDR - One and Two Stage 182

Chapter 1

Introduction - The Science of Cities

> And so a growing number of people have begun, gradually, to think of cities as problems in organized complexity – organisms that are replete with unexamined, but obviously intricately interconnected, and surely understandable, relationships. – Jane Jacobs *The Life and Death of Great American Cities*

More than half of the world's population now lives in urban areas[1]. The density of cities brings economic productivity, provides cultural amenities, and facilitates sustainability, but it is also the root of problems related to congestion, health, and safety. Cities are self-organized, complex adaptive systems and understanding their dynamics is critical for improving the lives of billions. With this goal in mind, a rich body of work has been devoted to understanding the particulars of urban life. Sociologists have detailed the countless daily interactions between neighbors on a city block[120]. Urban and transportation planners have explored how mobility, accessibility, and built form in a city can help direct its growth[143, 117, 67]. Urban economists have built models to understand productivity, prices, and human choice[26, 92, 240]. Despite great strides, progress in testing theories and hypotheses has been hindered by lack of data. The volume of interactions and behaviors to capture was too great and the expense of doing so left data a luxury only the richest cities could afford.

[1]United Nations Department of Economic and Social Affairs - World Urbanization Prospects - 2014 Update. `http://esa.un.org/unpd/wup/Highlights/WUP2014-Highlights.pdf`

The evolution of technology over the past decade has changed this. The rise of ubiquitous mobile computing allows billions of individuals to access people, goods, and services through smart devices such as cellular phones. The penetration of these devices is astounding. The six billion mobile phones currently in use triples the number of internet users and boast penetration rates above 100% in the developed word, e.g. 104% in the United States and 128% in Europe[2]. Even in developing countries, penetration rates are of 89%[3] and growing fast. These devices and the applications that run on them passively record the actions of their users including social behavior and information on location[4] with high spatial and temporal resolution. Cellular antennas, wifi access points, and GPS receivers are used to measure the geographic position of users to within a few hundred meters or less. While the collection, storage, and analysis of this data presents very real and important privacy concerns [71, 73], it also offers an unprecedented opportunity for researchers to quantify human behavior at large-scale.

The promise of new data and the need for new tools and techniques to make use of them has drawn an increasingly diverse set of domains to research on cities. Computer scientists have developed data collection, storage, and analytical techniques to monitor the interactions documented by urban ethnographers like Jane Jacobs and have built new services and applications to facilitate them. Physicists have identified scaling laws that allow them to predict many characteristics of a city based only its population [28, 29, 27] and modeled the dynamics of their growth using machinery from network science and statistical mechanics [20, 181, 22]. With billions of data points captured on millions of users each day, new research into computational social science [134] has begun to augment and sometimes replace sparse, traditional data sources, helping to answer old questions and raise new ones. Fueled by its great practical importance and new opportunities created by data availability, a new science

[2]GSMA European Mobile Industry Observatory 2011 `http://www.gsma.com/publicpolicy/wp-content/uploads/2012/04/emofullwebfinal.pdf`

[3]ITU. (2013) ICT Facts and Figures `http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2013-e.pdf`

[4]Lookout (2010) Introducing the App Genome Project `https://blog.lookout.com/blog/2010/07/27/introducing-the-app-genome-project/`

of cities is emerging. City Science is inherently interdisciplinary, infusing new methods and models into research on urban systems to ensure they are more sustainable, safer, and better places to live.

This thesis presents a series of projects that further our understanding of human behavior within cities. Chapter 2 begins by reviewing recent work that leverages novel and traditional data resources to understand human movement patterns. It describes a growing list of data sources that now capture human mobility and the infrastructure that enables it. While a large body of work has made use of data from surveys, this chapter focuses on more recent insights gleaned from applying methods from statistical physics and data mining to massive, passively collected data sets. These results are often more abstract that characterizations of mobility, but they provide an important foundation for turning new data sources into useful information. The remainder of this thesis builds on this foundation and more explicitly applies these insights to the study of economic, social, and transportation systems.

Chapter 3 is the first of these applications and presents empirical measurement and models of the relationship between mobility and social behavior in cities. It introduces two new metrics of mobility similarity between people and the predicability of an individuals visitation patterns. Measuring these metrics in three different cities reveals that individuals visit more of the same places with the same frequency as their social contacts than with strangers. Moreover, large portions of an individuals movements can be inferred by combining the movement patterns of these contacts. While social contacts are shown to inform mobility measurements, the reverse is also true. We show that measuring the mobility similarity at different hours of the day and week can reveal context about the nature of a social relationship (e.g. if two individuals are likely co-workers or close family members). Finally, this chapter presents the GeoSim model that reproduces these patterns far better than current methods. Algorithms that identify and enable opportunities to share are becoming increasingly important as the services from Uber to ZipCar allow individuals to connect and consume common transportation infrastructure.

In Chapter 4, we turn our attention to economic behavior. While new data sources

have proven their utility to improve forecasts of important indicators like disease, elections, or the stock market, surprisingly few applications exist using this data to understand economic behaviors. This chapter describes methods to detect mass layoffs of workers within a city, identify affected workers, and measures significant declines in social behavior and mobility following these events. We then show that measurements of these behaviors via mobile phone data can improve forecasts of unemployment rates at the province level within countries. These results not only shed light on behavioral changes that are prohibitively expensive to measure at scale, but demonstrate the use of new data sources as supplements to existing forecasts of critical economic indicators.

Chapter 5 presents a novel implementation of a four step model to estimate travel demand within cities from mobile phone data, open sourced maps, and census information. We discuss a number of algorithms to perform trip generation, trip distribution, mode choice, and route assignment using billions of mobile phone calls as input. We demonstrate the versatility of this model by estimating travel demand in multiple cities around the world. Particular care is taken to compare the output of this model with that of traditional four step models making use of travel survey data. We report high correlation between the two. This model implementation marks a major step towards making new data sources useful to urban and transportation planner by dramatically reducing the time between estimates and their cost.

In the same spirit of building new tools for urban planners, Chapter 6 describes applications of supervised learning methods to predicting land use from mobile phone activity within an area. While prediction errors are disappointingly high, a more careful examination of errors reveals inconsistencies with ground truth labels rather than the prediction techniques. These findings suggest that unsupervised learning methods may yield more informative zoning maps of cities, grouping places based on actual activity levels rather than potentially outdated and misleading regulations.

While the applications of these new data sources are important and far reaching, they also raise deep concerns about the privacy and wellbeing of the people who generate them. Chapter 7 discusses the increasingly important ethical and policy

ramifications of the collection and use of these personal data assets. By framing data as an economic asset, we explore the current market for data, its failures, and how various regulatory agencies may take actions to correct them. With billions of dollars in the balance, the answers to these questions will have a profound impact on technology for the foreseeable future.

Finally, we conclude with a summary of insights learned throughout this thesis and an outline for future work in the area.

Chapter 2

Human Mobility

Mobility has been a steering force for much of human history. The movement of peoples has determined the dynamics of numerous social and biological processes from tribal mixing and population genetics to the creation of nation-states and the very definition of our living areas and identities. Urban and transportation planners, for example, have long been interested in the flow of vehicles, pedestrians, or goods from place to place. Insights from models informed by novel data sources can identify critical points in road infrastructure, optimize public services such as busses or subways, or study how urban form influences its function and vice versa. Epidemiologists are also relying heavily on models of human movement to predict and prevent disease outbreaks [238, 64] as global air travel makes it possible for viruses to quickly jump continents and dense urban spaces facilitate human-to-human contagion. This has made understanding human movement a crucial part of controlling recent disease outbreaks [1]. Finally, social scientists are increasingly interested in understanding how mobility impacts a number of social processes such as how information spreads from person to person in offices and cafes across the world. These interactions have been theorized to impacts crime rates, social mobility, and economic growth [27, 171] and understanding their dynamics may improve how we live, work, and play.

The growing need to understand and model human mobility has driven a large body of research seeking to answer basic questions. However, the lack of reliable

[1] `http://www.worldpop.org.uk/ebola/`

and accessible data sources of individual mobility has greatly slowed progress testing and verifying these theories and models. Data on human mobility has thus far been collected through pen and paper surveys that are prohibitively expensive to administer and are plagued by small and potentially biased sample sizes. Digital surveys, though more convenient still require active participation and often rely on self-reporting [65]. Despite the development of statistical methods to carefully treat this data [25, 104, 170] new, cheaper, and larger data sources are needed to push our understanding of human mobility efforts further.

This chapter presents an overview of mobility research in the current data rich environment. It describes a variety of new data sources and detail the new models and analytic techniques they have inspired. We start by exploring research on individuals that emphasizes important intrinsic and universal characteristics about our movement: we are slow to explore, we are relatively predictable, and we are mostly unique. We then discuss efforts to add context and semantic meaning to these movements. Finally, we review research that models aggregates of human movements such as the flow of people from place to place. This chapter also serves to highlight applications of this research to areas such as congestion management, economic growth, or the spreading of both information and disease.

2.1 New Data Sources

Traditional data sources for human mobility range from census estimates of daily commutes to travel diaries filled out by individuals. These surveys are generally expensive to administer and participate in as they require intensive manual data encoding. To extract high-resolution data, individuals are often asked to recall large amounts of information on when, where, and how they have traveled making them prone to mistakes and biases. These challenges make it hard for surveys to cover more than a day or week at a time or to include more than a small portion of the population (typically less than 1%).

Mobile phones, however, with their high penetration rates, are an extremely useful

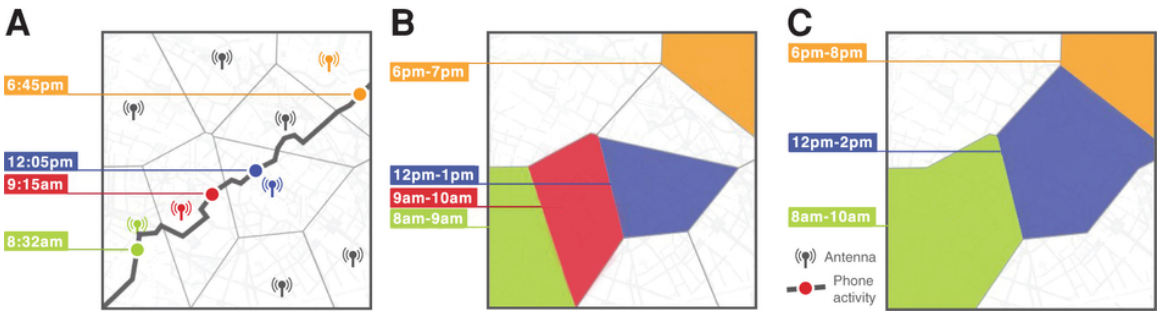

Figure 2-1: Mobile phones are increasingly being used to collect high-resolution mobility data. This figure from de Montjoye et al. [71] depicts A) a sequence of calling events made by a user at different locations. B) These events are localized to the area served by the closest mobile phone tower to the use and C) can be aggregated into individual specific neighborhoods where a user is likely to be found at different times of the day or week.

sensor for human behavior. A large fraction of location data from mobile phones are currently in the form of call detail records (CDRs) collected by carriers when users perform actions on their devices that make use of the telecommunications network. The location of each device at the time a call, text, or data request is registered (Figure 1) is recorded by carriers for billing, network performance, and legal purposes. Locations are inferred either by observing the tower through which the phone is connected or by triangulation with nearby towers. With the increasing use of mobile phones, each individual generates tens to hundreds of these digital breadcrumbs on a daily basis and this number is only increasing. Through specific agreements or through open-data challenges [74], location data on millions of users is readily available to researchers and has been used extensively to augment and sometimes replace traditional travel surveys. This data now forms the core of numerous new mobility studies and models some of which we describe below.

Though generally less common than CDRs, applications running on smartphones may access even more precise estimates of a user's position. A variety of these sensors, from GPS to wifi, can pinpoint the location of a device to within just a few meters and can record data every few minutes [5]. Similarly, protocols such bluetooth and NFC allow devices to discover and connect to one another within a few meter radius, creating ad hoc sensor and social proximity networks [81]. Some of these applications and underlying social-networks explicitly add crucial context to

mobility data. Foursquare invites users to "check-in' at specific places and establishments, Twitter will automatically geotag tweets with precise coordinates from where they were sent, and the Future Mobility survey app passively maintains an activity diary [65] requiring little input from users.

Infrastructure and public services have also become much smarter and now collect data on their usage to improve and help plan operations. Toll booths automatically count and track cars and this data has helped create accurate and real-time traffic estimates used by mapping and navigation services to provide better routing information. Subways, streetcars, and busses use electronic fare systems that record when millions of users enter and exit transportation systems to help better predict demand. In addition to smarter public infrastructure, the ecosystem created by digital devices has given birth to entirely new transportation services such as Hubway, the Boston bike rental service, that collects data on every bike ride and has even released some publicly [2] or Uber, an on-demand car service, that uses historical usage data to balance the time a user has to wait for a car to arrive and the time drivers spend without clients. Finally, on-board devices and real-time data feeds from automatic vehicle location (AVL) systems power applications such as NextBus to track the location of thousands of busses and subways across the world to display and predict when the next bus will arrive. While smart infrastructure comes with its own privacy challenges [126] [3], vehicle and public transport data offer additional information to urban planners and mobility modelers to better understand these systems.

Finally, most practical mobility models need to properly account for geography such as mountains and rivers, transportation infrastructure such as bridges and highways, differences in density between urban and rural areas, and numerous other factors. Thankfully, the digitization of maps has led to an explosion of geographic data layers. Geographic information systems (GIS) have improved dramatically while falling data storage prices have made it possible for small and large cities to offer

[2]Hubway Data Visualization Challenge (2012) `http://hubwaydatachallenge.org/`

[3]New York taxi details can be extracted from anonymized data, researchers say (2014) `http://www.theguardian.com/technology/2014/jun/27/new-york-taxi-details-anonymised-data-researchers-warn`

their public mapping data to citizens in an online, machine readable format. The U.S. Census Bureau's TIGERline program, San Francisco's OpenSF, and New York City's PLUTO data warehouse are just a few sources that offer huge repositories of publicly accessible geographic data on everything from building footprints and the location of individual trees in a city. Open- and crowd-sourced initiatives like OpenStreetMap allow anyone in the world to contribute and download high-resolution digital maps of roads, buildings, subways, and more, even in developing areas that may not have institutional resources to create them. Private efforts such as Google Maps and MapBox offer high-resolution satellite imagery, route planning, or point of interest information through free or low cost APIs. Put together, these resources provide a digital map of the world that serves as a rich backdrop on which to study human mobility and the infrastructure built to facilitate it.

Put together, new sources from CDRs to public transport data, from mobile phone applications to AVLs, generate a dataset with size and richness prohibitively expensive to match via traditional methods. Collected passively and without any effort from the user, this data is often more robust to manipulation by conscious or unconscious biases and provide a signal that is difficult to fake. While we are convinced of the potential of this data, it is always important to remember that it is not without pitfalls. It would be illusory to think that all of the old biases or hidden variables would simply disappear because the data is large. In some cases, data is only recorded when an individual interacts with a device which may bias when samples are taken [173]. Similarly it is important to keep in mind that even if it covers a significant fraction of the population this data might not be representative. Finally, these data generally come stripped of context. We do not know why an individual has chosen to move or what they will be doing there. For these reasons, sampling and robust statistical methods are still–maybe more than ever [4]–needed to use this data to augment our current understanding of human mobility while still providing robust conclusions. We now discuss a number of studies that aim to do just this.

[4]Flowing data - Where People Run in Major Cities `http://flowingdata.com/2014/02/05/where-people-run/`

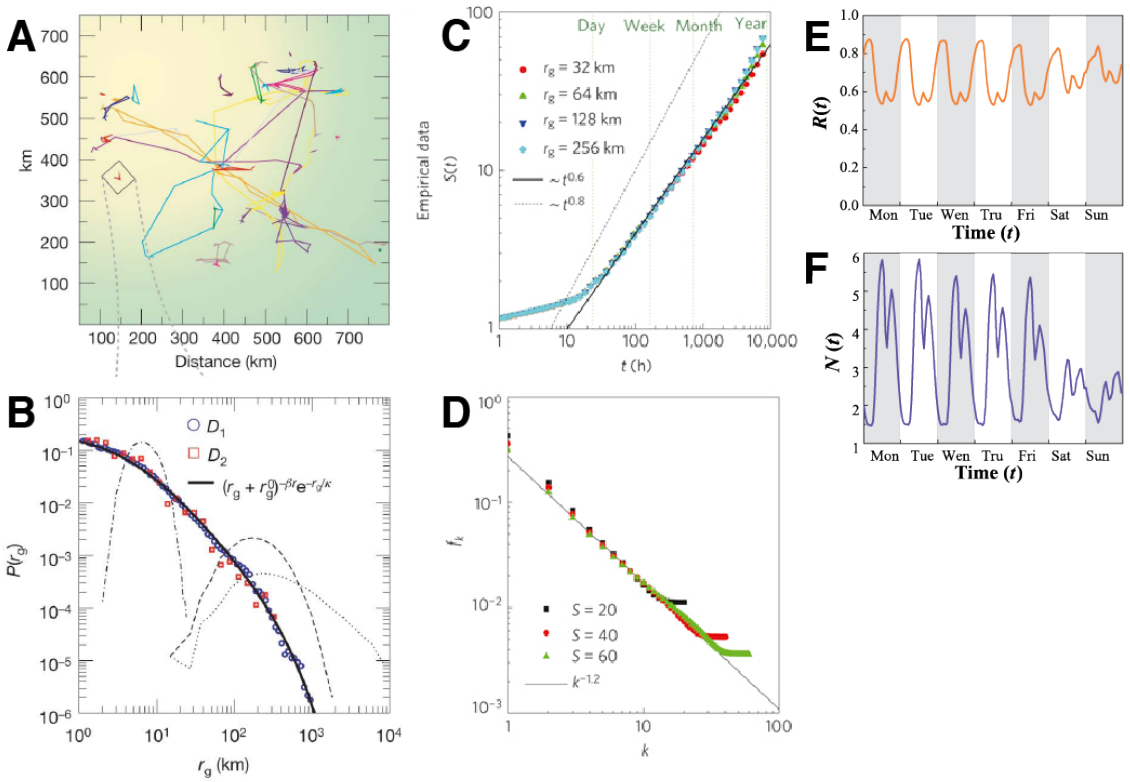

Figure 2-2: A) Individual mobility trajectories are passively collected from mobile devices [97]. B) Measuring the distribution of radius of gyrations, r_g within a population of 100,000 users in a European country reveals considerable heterogeneity in typical travel distance of individuals. Moreover, this distribution cannot be explained by modeling each individual's movement as realizations of a single Levy flight process [97]. C and D) Show the slower than linear growth in new locations visited over time $S(t)$ and that the probability a location is visited next is inversely proportional to the frequency it has been visited in the past [202]. E) This preferential return contributes to strikingly high predictability $R(t)$ over time while F) the number of unique locations visited in any given hour is highly periodic and corresponds to the sleep-wake cycles of individuals [205].

2.2 Individual Mobility Models

Understanding mobility at an individual level entails collecting and analyzing sets of times, places, and semantic attributes about how and why users travel between them. For example, on a typical morning one may wake up at home, walk to a local coffee shop on the way to the bus that takes them to work. After work they may go to the grocery store or meet a friend for dinner before returning home only to repeat the process the next day. The goal modeling this mobility is to understand the underlying patterns of individuals using new high resolution data. While these models have been

used to plan infrastructure or public transport they have also uncovered insights into the underlying nature of human behavior: we are slow to explore, relatively predictable, and mostly unique.

Early modeling work draws a great amount of inspiration from statistical physics, with numerous efforts making parallels with human mobility and random walk or diffusion processes. One of the used data from the crowdsourced 'Where's George' project. Named after George Washington, whose head appears on the $1 bill, the project stamped bills asking volunteers to enter the geographic location and serial number of the bills in order to build a travel history of various banknotes. As bills are primarily carried by people when traveling from store to store, a note's movement serves as a proxy for human movement. Modeling the bills trajectories as continuous random walks, Brockmann et al. found that their movement appears to follow a Levy flight process [36]. This process is characterized by subsequent steps whose angular direction is uniformly distributed, but whose step-lengths follow a fat-tailed distribution. While small jumps are most probable, bills have a significant probability of making long jumps from time to time. These findings are aligned with observations that humans tend to make many short trips in a familiar area, but also take longer journey's now and then.

In 2008, Gonzalez et al. [97] showed that the movement of these bills does not tell the whole story. Using a CDRs dataset of more than 100,000 users over a 6 month period in a European country (Figure 2A), they showed that the step-length distribution for the entire population was better approximated by a truncated power-law $P(\Delta r) = (\Delta r + \Delta r_0)^{-\beta} \exp(-\Delta r/\kappa)$ with exponent $\beta = 1.79$ and cutoff distances between 80 km and 400 km. This suggests that Levy flights are only a good approximation of individual's mobility for short distances. To understand the mechanism that gives rise to this distribution, the authors borrowed a quantity from polymer physics known as the "radius of gyration" r_g:

$$r_g(t) = \sqrt{\frac{1}{N(t)} \sum_{i=1}^{N(t)} (\mathbf{r} - \mathbf{r_{cm}})^2}, \tag{2.1}$$

where N(t) are the number of observed locations and r_{cm} is the mean location of the user during the observation period. In essence, the radius of gyration characterizes the radius of the circle an individual can most often be found in during an observation period t. The authors then showed that the distribution of r_g in the population is itself well approximated by a truncated power-law with $r_g^0 = 5.8$km, $\beta_{r_g} = 1.65$, and a cutoff of $\kappa = 350$km (Figure 2B). Simulations suggest that the step-length distribution of the entire population is produced by the convolution of heterogeneous Levy flight processes, each with a different characteristic jump size determined by an individual's radius of gyration. Put differently, each person's mobility can be approximated by a Levy flight process up to trips of some individual characteristic distance r_g. After this distance, however, the probability of long trips drops far faster than would be expected from a traditional Levy flight.

Further investigation by the authors revealed the source of this behavior: the idiosyncrasy of human movements. Unlike random processes, humans are creature of habits and tend to returns to previously visited locations such as home or work. The nature of these returns was also found to follow a very particular pattern. An individual returns to a previously visited location with a probability proportional to that location's rank $P(L)$ $1/L$ amongst all the places he or she visits. These non-random, predictable return visits are unaccounted for in random walk and Levy flight models and have been shown to be at the heart of deviations of observed behavior from random processes. Additional studies [54] have found similar patterns in both other CDRs datasets and Foursquare or Twitter check-ins.

Subsequent work by Song et al. [202] further studied how individual-specific locations need to be taken into account in mobility models. Using a similar CDR dataset, the authors showed three important characteristics of human behavior. First, the number of unique locations visited by individuals $S(t)$ scales sub-linearly with time $S(t)$ t^{μ} where $\mu = 0.6$ (Figure 2C). Second, the probability an individual returning to a previously visited locations scales with the inverse of the rank of that location $P(L)$ $L^{-\zeta}$ where $\zeta = 1.2$ (Figure 2D), a phenomena labeled as 'preferential return'. And third, the mean displacement (Δr) of an individual from a given starting point

shows slower than logarithmic growth, demonstrating the extremely slow diffusion of humans in space. In essence, these finding pinpoint the dampening of explorative human movement overtime. Long jumps are observed so infrequently that they do not affect the average displacement of individuals. The authors then propose a new model of human mobility to capture these three characteristics. Starting at time t, an individual will make a trip at some future time Δt drawn from a fat-tailed probability distribution measured from CDRs. With probability $\rho S^{-\gamma}$, the individual travels to a new, never-before visited location some distance Δr away, where Δr is drawn from the fat-tailed distribution characterized in the previous model. With probability $1-\rho S^{-\gamma}$ an individual returns to a previously visited location according to the inverse rank equation.

These early models do not attempt to recover periodic aspects of movement (e.g. daily commuting) or semantic meaning of visits (e.g. to visit a friend or go shopping), or attempt to do so. They do, however, emphasize important statistical and scaling properties of human mobility and often successfully reproduce them. Taken together, these models show humans are slow in our exploration, returning more often than not to known places and with less long steps than predicted by a power-law distribution.

Approaching the problem from the perspective of machine and statistical learning, another set of models has uncovered and explored another facet of human mobility: how predictable we are. In [205], Song et al. used information theory metrics on CDRs to show the theoretical upper-bound on predictability using three entropy measures the entropy S, the random entropy S^{rand}, and the uncorrelated entropy S^{unc}. They then use their empirical distributions to derived an upper bound on a user's predictability ($\prod^{max}$, $\prod^{rand}$, and $\prod^{unc}$). On average, the potential predictability of an individual's movement is an astounding 93% and no user displayed a potential predictability of less than 80%. To further quantify predictability, the author introduced two new metrics. They defined regularity $R(t)$ as the probability a user is found at their most visited location during a given hour t, along with the the number of unique locations visited during a typical hour of the week $N(t)$ (Figure 2E and 2F). Both show strong periodicity and regularity. These quantities have since been measured

in different data sets in different cities and countries and have been shown to be consistent among them [54].

While the previous study provided a theoretic upper bound on the predictability of an individual, a number of statistical learning techniques have been developed to make predictions of where an individual will be at a given time. Early work in the area, predating even analytic computations, used Markov models and information on underlying transportation networks to predict transitions between mobile phone towers within cities. These models have been used to improve quality of service of wireless networks through proper resource allocation [124, 137, 218, 131]. Later work incorporated various trajectory estimation and Kalman filtering algorithms to predict movements in small spaces such as college campuses [157, 135].

Temporal periodicity was used by Cho et al. [58] in their Periodic Mobility Model and social behavior incorporated in the Period Social and Mobility Model. At their core, these models are mixture models in two-dimensional space that learn the probability distribution of a user to be at any given location at a given time from previous location data. The latter also account for the location history of social contacts. The authors used these models to estimate that as much as 30% of our trips may be taken for social purposes. Multivariate nonlinear time series forecasting produced similar results [78, 188] predicting where an individual will be either in the next few hours or at a given time of a typical day. These models, however, are all focused on predicting the geographic position of individuals at different times and do not attempt to understand what individuals may be doing there or any other semantics of place.

Though acquiring semantic information about mobility is more difficult than simply measuring geographic coordinates, it provides a much richer abstraction to study behavior. In one of the first studies to mine the behavior of college students using mobile phones, Eagle et al. [80] gave a few hundred students smart phones that recorded not only locations, but asked users to label each place with its function such as home or work. Applying principal component analysis to these abstract movements from semantic place to semantic place (as opposed to geographic movements alone), the authors found that an individual's behavior could be represented as a linear combina-

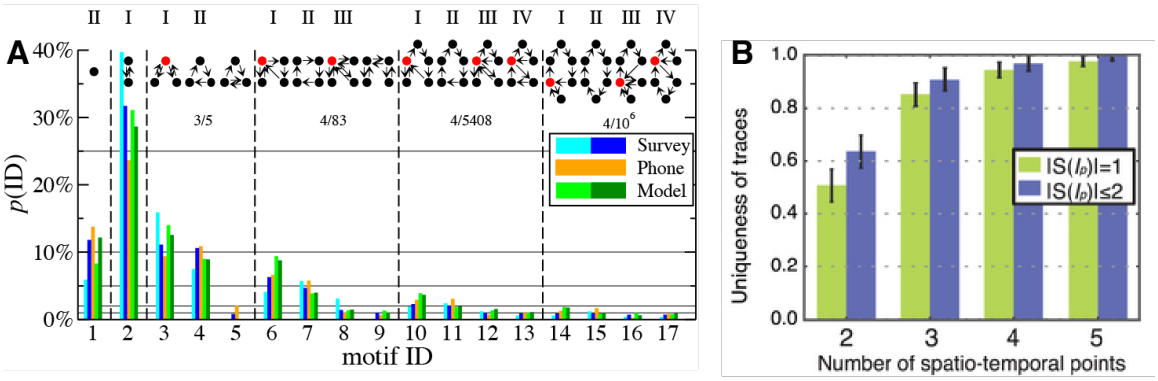

Figure 2-3: A) Removing geographic coordinates from locations and only focusing on a set of unique places and the directed travel between them, mobility motifs reveal that the daily routines of people are remarkably similar. Despite over 1 million unique ways to travel between 6 or fewer points, just 17 motifs are used by 90% of the population. Moreover, the frequency of their appearance in CDR data matches very closely with more traditional survey methods [190]. B) Despite this similarity and predictability, our movement displays a high degree of unicity. Just four spatiotemporal points is enough to differentiate a user from 95% of all others individuals [71].

tion of just a few 'eigenbehaviors'. These eigenbehaviors are temporal vectors whose components represent activities such as being at home or being at work. They can be used to predict future behaviors, perform long range forecasts of mobility, and label social interactions [81, 184]. The price paid for such detailed predictions, however, is the need for semantic information about locations. Geographic positions need to be tagged with attributes such as home or work in order for them to be grouped and compared across individuals.

Another approach to studying more abstract measurements of individual location information comes from recent work by Schneider et al. [190]. The authors introduced *mobility motifs* by examining abstract trip chains over the course of a day. A daily mobility motif is defined a set of locations and a particular order that a person visits them over the course of a day. More formally, these motifs constitute directed networks where nodes are locations and edges are trips from one location to another. For example, the motif of an individual whose only trips in a day are to and from work will consist of two nodes with a two directed edges (one in both directions). Counting motifs in mobility data from both CDRs and traditional travel surveys, they find on average individuals visit 3 different places in a given day. They then

construct all possible daily motifs for a given number of locations n and compute the frequencies that those motifs appear in human mobility data. Shockingly, while there exist over 1 million ways for a user to travel between 6 or fewer locations, 90% of people use one of just 17 motifs and nearly a quarter follow the simple two location commute motif introduced earlier (Figure 3A). The authors found similar results in travel survey data and introduced a simple Markov model for daily mobility patterns which reproduces empirical results.

It is tempting to hypothesize that high theoretical and practical predictability results from high levels of similarity between individuals in a region. Perhaps the pace of life, pull of mono-centric downtowns, or the structure of transportation systems funnel users to the same places and route choices. de Montjoye et al. [71] explored this hypothesis and found that, while predictable, an individual's movement patterns are also unique. The authors introduced *unicity*, $\mathcal{E}_p$, as the fraction of traces uniquely defined by a random set of p spatiotemporal points where a trace T is a set of spatiotemporal points, each containing a location and a timestamp. A trace is said to be uniquely defined by a set of points I_p if it is the only trace that matches I_p in the entire dataset. Applying this measure to a CDR dataset on 1.5 million users, the authors found that just four spatiotemporal points is enough to uniquely identify 95% of all users (Figure 3B). The authors further study unicity when the data is coarsened spatially or temporally. They found $\mathcal{E} \sim (v * h)^{\beta}$ unicity decrease as a power function with the spatial (v) and temporal resolution of the data (h) and that $\beta \sim -p/100$. Taken together, these equations show that unicity decreases slowly with the spatial and temporal resolution of the data and that this decrease is easily compensated by the number of points p. High uniqueness in human mobility traces exists across many spatiotemporal scales. These results raise many questions about the privacy of massive, passively collected metadata datasets, but also highlight an interesting nuance of human mobility: though individuals are predictable, they are also unique.

Merging concepts of predictability and unicity, work by Sun et al. [217] used temporal encounter networks to study repeated co-locations between passengers using data from bus passengers in Singapore. Temporal encounter networks were con-

structed by connecting individuals if they rode the same bus at the same time. An average individual encountered roughly 50 people per trip and these trips were highly periodic, occurring at intervals associated with working hours as well as daily and weekly trips. A pair of individuals who encountered each other tended to meet an average of 2.5 times over the course of a week. The distribution of time between encounters reveals strong periodicity, with passengers riding the same bus to work in the morning riding the same home, or riding the same bus at the same time each morning. This finding illustrates the idiosyncrasies of human mobility. We not only visit just a few places during the day, we do so at the same times and by the same routes. Though both of these results suggest that our unicity should be low, the previous work shows us that this is not the case.

In summary, new data sources have allowed researchers to show that, over weeks and months, human movement is characterized by slow exploration, preferential return to previous visited places, exploration of daily motifs, and predictable uniqueness. These regularities have been used to develop algorithms capable of predicting movement with high degrees of accuracy and have been shown to mediate other important processes such as social behavior and disease spread. Individual mobility patterns, however, are not the only level of granularity of interest to researchers, city planners, or epidemiologist. Aggregate movement can be either derived from individual level model or modeled as an emergent, personified phenomena. In the next section, we discuss works and models which aim at describing and modeling aggregate movement and flows of many individuals from place to place.

2.3 Aggregate Mobility

Aggregated mobility is used for planning urban spaces, optimizing transportation networks, studying the spread of ideas or disease, and much more. Perhaps the largest component in these models are origin-destination matrices that store the number of people traveling from any location to any other at different times or by different means. Like many complex systems, aggregate behavior is often more than the sum

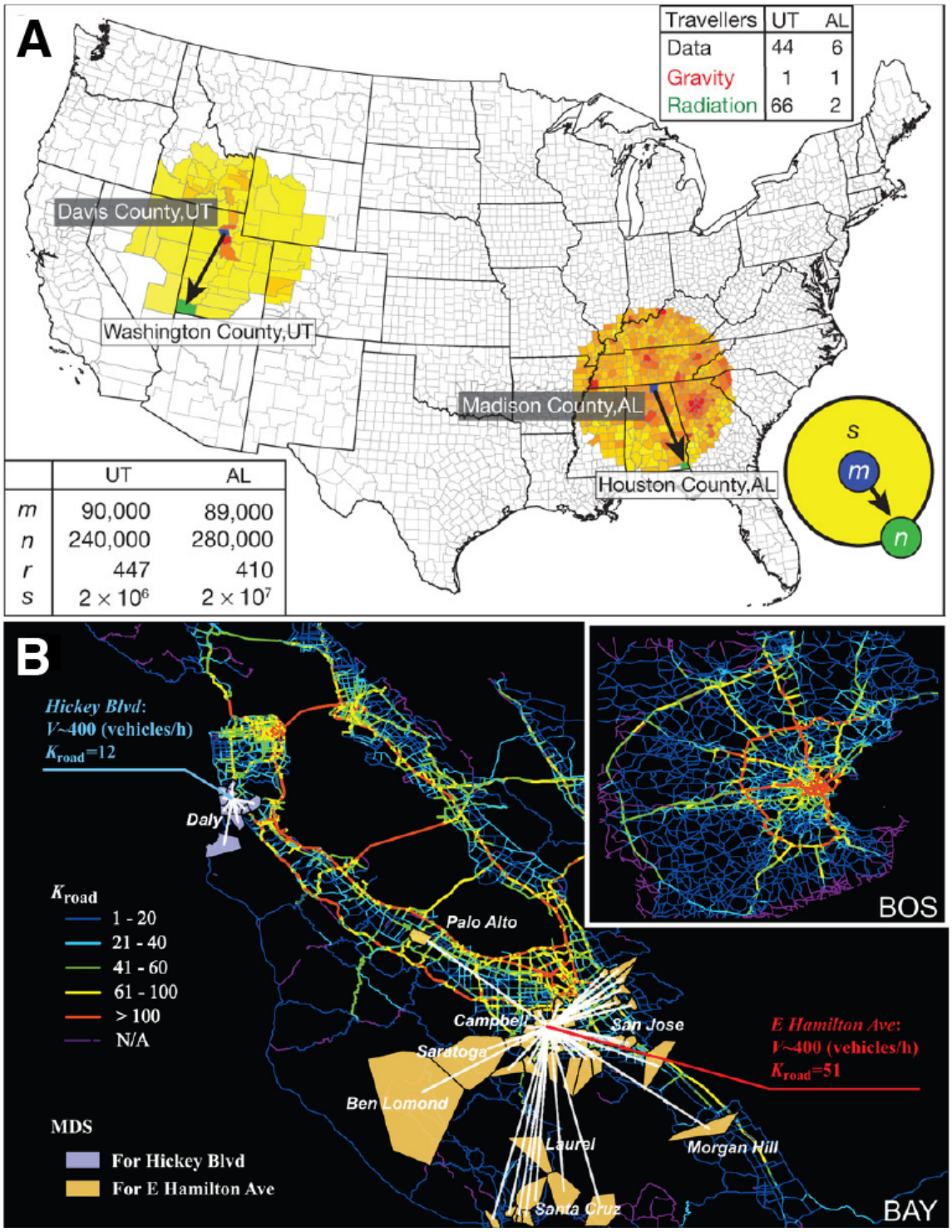

Figure 2-4: A) The radiation model accounts for intervening opportunities, producing more accurate estimates of flows between two places than more traditional gravity models [200]. B) Routing millions of trips measured from CDR data to real road networks makes it possible to measure the importance of a road based on how many different locations contribute traffic to it, K_{road}. Understanding how transportation systems perform under different loads presents new opportunities to solve problems related to congestion and make infrastructure more efficient [229].

of individual parts and can be modeled separately. Additional layers of complexity are also needed to account for and sometimes explain individual choice of mode of transportation or route as described by the “four step model" [150, 170].

Like their individual-focused counterparts, many of these aggregate models are inspired by physical processes. Some of the earliest techniques for estimating origin-destination matrices are gravity models which have been used to model flows on multiple scales, from intra-city to international [106, 170]. Borrowed directly from Newton's law of gravitation, the number of trips T_{ij} taken from place i to place j is modeled as a function of the population of each place m_i and m_j and some function of the the distance between them $f(r_{ij})$. The intuition is that the population of a place, it's mass, is responsible for generating and attracting trips and thus the total flux between the two places should be proportional to the product of the two masses while the distance between them mitigates the strength of this connection. In the fully parameterized version of this model, an exponent is applied to the population at the origin and destination $T_{ij} = a\frac{m_i^\alpha m_j^\beta}{f(r_{ij})}$ to account for hidden variables that may be specific to local regions or populations. While the classical gravity model from physics is recovered by setting $\alpha = \beta = 1$, and $f(r_{ij}) = r_{ij}^2$, these parameters are generally calibrated for specific application using survey data.

Gravity models, however, are not without limitation. First, they rely on a large number of parameters to be estimated from sparse survey data which often leads to overfitting and, second, they fail to account for opportunities that exist between the two masses of people. The latter fault results in the same flow of people being estimated between two locations whether there is an entire city or an empty desert between them. Intuitively, one would expect that trips between places would be affected by the intervening opportunities to complete a journey. These shortcomings led Simini et al. to develop the radiation model [200]. Again borrowing from physics (this time radiation and absorption), they imagined individuals being emitted from a place at a rate proportional to its population and absorbed by other locations at a rate proportional to the population there. In this model, the probability that an emitted person arrives at any particular place is a function of their probability of not being

absorbed before getting there. The model is as follows: $T_{ij} = T_i \frac{m_i m_j}{(m_i+s_{ij})(m_i+m_j+s_{ij})}$, where T_i is total number of trips originating from location i and s_{ij} is the population within a disc centered on location i with a radius equal to the distance between i and j. The radiation model does not directly depend on the distance between the two places, taking instead into account the opportunities in-between them (Figure 4A). Unlike the gravity model, the radiation model is parameterless and requires only data on populations to estimate flow. The authors showed that despite its lack of parameters, the radiation model provides better estimates of origin-destination flows than the gravity model for areas the size of counties or larger.

Yang et al. adapted Simini's radiation model to correct for distortions caused at different scales [242]. They showed the original radiation model's lower accuracy in urban environment is due to the relatively uniform density and small distances that characterize cities. In dense urban areas, distances are all relatively short and an individual may choose to visit a particular location due to hedonic attributes regardless of whether it is convenient to get to or not. Yang et al. subsequently introduced a scaling parameter α in the function describing the conditional probability an individual is absorbed at a location. This single parameter was enough to correct for these distortions and to provide a model that works on any length scale. Moreover, the authors suggested that for urban areas, the density of points of interest (POIs) such as restaurants and businesses is a better predictor of the absorption of a place than its population. Iqbal et al. [119] have demonstrated an improved way to extract valid, empirical OD matrices from call detail records (CDRs) data to validate the model.

Finally, activity-based models [25] model user intent more explicitly. They hypothesize that all trips are made to fulfill certain needs or desires of an individual. Travel and survey diaries are used to identify those needs for different segments of the population and how they are typically fulfilled. This knowledge can then be used by the model given the demographics of individuals and environmental factors. These models are closely related to agent-based models simulating the behavior of city residents and rely heavily on the idea of economic utility. These models are often used

to model what mode of transportation a person will use to get from place to place. While private automobiles are a popular choice in places like the United States, they are far from the only choice, especially within cities. Public transportation, bicycles, and walking are all plausible alternatives. Digital sensors like smartphones or automatic vehicle location (AVL) systems are increasingly used by researchers to asses transportation system performance and even map informal systems like those found in many developing cities[77, 57, 187].

As human movement via cars, trains, or planes has always been a major vector in the propagation of diseases, epidemiologists are increasingly turning to mobility research to inform their models of disease spread. For example, CDR data has been used to map mobility patterns in Kenya helping researchers in their fight against Malaria [238, 237]. More recently, CDR and other data from West-Africa has been used to model regional transportation patterns to help control the spread of Ebola [5]. Finally, air travel data has become central to the study of global epidemics when planes allow an individual to travel between nearly any two points on the globe in a matter of hours. The global airline network therefore often determines how potent an epidemic could be and its likely path across the globe [163, 15, 16, 151, 64] (Figure 5A).

In this chapter, we reviewed a number of ways new data sources are expanding our understanding of human mobility. Applying methods from statistical physics, machine learning, and traditional transportation modeling, reproducible characteristics of human movement become visible. We explore slowly [97, 202], we are highly predictability [205, 78], and we are unique [71]. Models of aggregate flows of people from place to place have also found success with analogies to statistical physics validated by new data sources [200]. New sensors and data collection systems promise better monitoring and operation of transportation systems from cars to busses and aircraft. Valuable in their own rights, these insights have informed our understanding of other social phenomena as well, leading to more accurate models of disease spread,

[5] Cell-Phone Data Might Help Predict Ebola's Spread (2014) `http://www.technologyreview.com/news/530296/cell-phone-data-might-help-predict-ebolas-spread/`

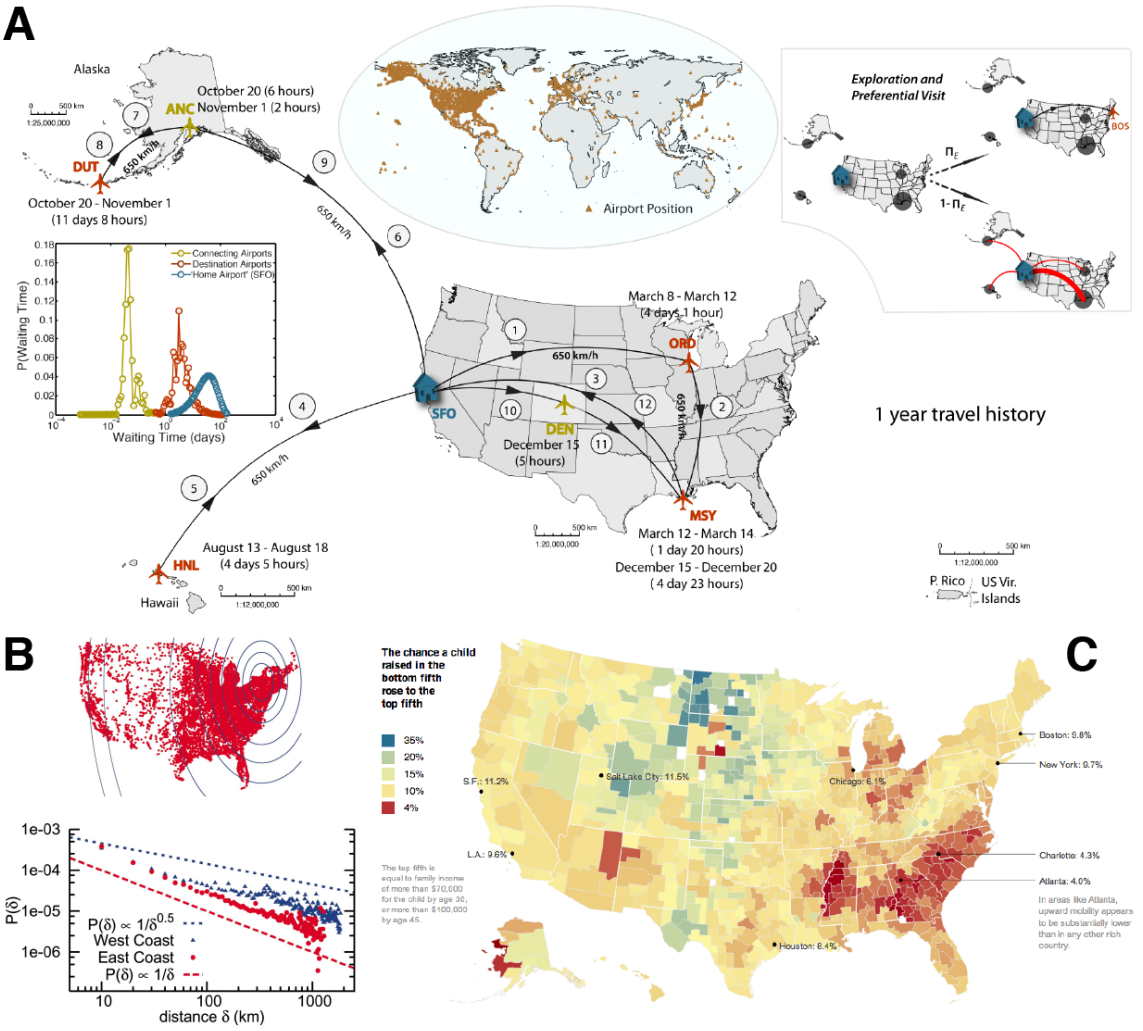

Figure 2-5: A) Global air travel has dramatically increased the speed at which diseases can spread from city to city and continent to continent [163]. B) Mobility also impacts social behavior as we are far more likely to be friends with someone who lives nearby than far away [136]. C) Mobility and the access it provides has strong correlations with economic outcomes. Children have dramatically different chances at upward economic mobility in certain places of the United States than others [55].

social interactions, and economic outcomes. As cities become home to millions more people each year, the insights gained from these new data are critical for making them more sustainable, safer, and better places to live.

2.4 Human Mobility and City Science

While human mobility is of obvious interest to travelers, urban planners and transportation engineers, people's movement strongly impacts other areas. The remainder of this thesis explores the relationship between human mobility and three other aspects of city life: social behavior and choice, economic behavior within cities, and the use and performance of city infrastructure.

2.4.1 Social Behavior and Choice

Choice is a crucial element of human mobility and movement is often a means to a social end. Despite new communications technologies making it easier than ever to connect across vast distances, face to face interactions still play an important role in social behavior whether it is the employees of a company commuting to a central workplaces or friends meeting at a restaurant on a weekend. The link between social contacts and mobility choices has become increasingly prominent in research as mobility data is often collected through mobile phones or location-based social networks.

Using data from an online social-network, Liben-Nowell showed the probability of being friends with another individual to decrease at a rate inversely proportional to the distance between them suggesting a gravity model of the form discussed above [136] (Figure 5B). Subsequent work verified Liben-Nowell findings in other social networks [12, 99] while Toole et al. [220] showed the importance of taking into account geography when studying social-networks and how information spreads through them. Moreover, geographic characteristics can be used to predict the social fluxes between places [115]. Conversely, social contacts are very useful in predicting where an individual would travel next [58, 78, 222] and Cho et al. find that while

50%-70% of mobility can be explained as periodic behavior, another 10%-30% are related to social interactions.

Models such as the one proposed by Grabowicz et al. [99] have subsequently been developed to incorporate this dynamic and evolve both social networks and mobility simultaneously. The authors incorporate social interactions by having individuals travel in a continuous 2D space where an individual travel's is determined by the location of their contacts and use location as a determinant of new social tie creation. The model is as follows: with probability p_v, an individual moves to the location of a friend, and, with probability $1 - p_v$, they choose a random point to visited some distance Δr away. But, while social ties impact mobility, mobility can also impact social ties. Upon arriving at a new location, the individual can thus choose to form social ties with other individuals within a radius with probability p or random individuals anywhere in the space with probability p_c, a free parameter. A simple model is here also able to reproduce many empirical relationships found in social and mobility data.

These studies re-enforce knowledge that social forces are strong motivators. They influence decisions and choices people make on where to go, what to buy, and what to believe. Chapter 3 describes empirical methods to quantify similarity in movement patterns between social contacts and measure correlations between social behavior and mobility. Results show that temporal variations in mobility similarity can be used to classify social relationships and present an extension to a popular mobility model that captures these empirical findings. There is hope that these findings may help improve forecasts of travel demand within cities, add much needed context on the nature of relationships to large social networks, and enable the next generation of social applications that promote real change in people's behaviors and experiences for the better.

2.4.2 Economics Behavior in Cities

Mobility not only provides people with social opportunities, it also provides economic ones. Economists and other social scientists have developed numerous theories on the

role of face to face interactions in socio-economic outcomes and economic growth. In-person meetings are thought to unlock human capital, making us productive [125, 88]. For example, jobs in dense cities tend to pay higher wages than the same jobs in more rural areas even after controlling for factors such as age and education [243] in part due to productivity and creativity gains made possible by the rich face to face interactions that close spatial proximity facilitates. Universal urban scaling laws have been repeatedly found showing that societal attributes from the number of patents to average walking speed scales with population and theoretic models have been proposed that suggest density is at the heart of these relationships [28, 27, 171]. While density is one way to propagate these benefits, increased mobility is another. Poorer residents of cities have for example been shown to have better job prospects and higher chances of retaining jobs when given a personal car instead of being constrained by public transit [101]. Finally, Chetty et al. [55] found strong correlations between intergenerational economic mobility and variables related to the commuting times and spatial segregation of people (Figure 5C). While we are only beginning to explore these relationships, early returns suggest that mobility is a critical component of many economic systems.

To this end, Chapter 4 describes the use of mobile phone data to track employment shocks. At the micro level, it is demonstrate detecting a mass layoff in a town and identifying affected individuals by observing changes in calling behavior. Mobile phones are exceptional sensors for measuring the impact of layoffs on individuals, showing persistent drops in a user's social behavior and mobility. Finally, this work show that we can improve forecasts of unemployment rates months before traditional surveys are conducted and released by incorporating measured changes in social and mobility behavior in populations of users into predictions of unemployment rates at the regional level. These results suggest the potential of this data to improve measurements of critical economic indicators.

2.5 Mobility and City Infrastructure

From a practical perspective, city planners need to know not only how many people will go from point A to point B at a certain time of the day but also the mode of transportation and route choice of these individuals. For example, we would like to predict which route they will take so that we can properly estimate the stress placed on transportation systems and potentially optimize performance. Models of route choice typically assume that individual rationally chose the path from A to B that minimize some cost function such as total travel time or distance. Paths can be computed on a road network using shortest path algorithms such as the traditional Dijkstra algorithm or A-Star, an extension that enjoy better performance thanks to heuristics. Other informations such as speed limits can also be taken into account to estimate free flow travel times.

More advanced models are needed to account for the impact of congestion as drivers rarely encounter completely empty freeways. Incremental traffic assignment algorithms model congestion endogenously [211]. Trips are first split into increments containing only a fraction of total flow between two points. Trips in each increment are then routed along shortest paths independently of all other trips in that increment keeping counts of how many trips were assigned to each road. The travel times are then adjusted according to a volume delay function that accounts for the current congestion on a road where congestion is computed as the ratio between the volume of traffic assigned to the segment and the capacity of the road (referred to as volume-over-capacity). Trips the next increment are then routed using updated costs until all flow has been accounted for. In this way, as roads become more congested and the travel time increases, drivers in later iterations are assigned to different, less congested routes. Values of total volume on each road, congestion, and travel times can then be validated against traffic counters, speed sensors, or data from vehicle fleets like taxis and busses but also smartphones such as in the Mobile Millennium project [186, 116, 114, 121].

Wang et al. [229] further explored the use of CDRs as input for these iterative

algorithms to estimate traffic volume and congestion. After correcting for differences in market share and vehicle usage rates, they measure trips by counting consecutive phone calls of individuals as they move through the city to generate flow estimates that were then routed. Using this approach, Wang et al. show the distribution of traffic volume and congestion to be well approximated by an exponential mixture model. This model depends on the number of major and minor roadways in a cities network. Using the same approach, the authors describe the usage patterns of drivers by a bipartite usage graph connecting locations in the city to roads used by those travelers (Figure 4B). Roads can be defined by the number of locations that contribute traffic them and places can be described by the roads used to visit. The "function' of a road can then be classified by comparing its topological to its behavioral importance. For example, a bridge may be topologically important because it is the only way to cross a river, but a main street may be behaviorally important because it attracts motorists from many different neighborhoods. Using these measures, researchers were able to devise congestion reduction strategies that target the 2% of neighborhoods where trip reduction will have the largest network wide effect. They found this smart reduction strategy is three to six times as effective as a random trip reduction strategy. Further work used this analysis to predict traffic jams [228, 231].

Chapter 5 describes a software system that generalizes these results and provides analysis of road usage patterns in many cities across the world. This platform is used to conduct comparative studies of congestion patterns in cities and explore the relationship between network topology, system performance, and travel demand.

2.6 Acknowledgements

The work in this chapter was the result of a collaboration with Yves-Alexandre de Montjoye, Marta C. González, and Alex Pentland.

Chapter 3

Social Behavior and Choice

3.1 Introduction

Each day people make choices about where to go, how to get there, and who to go with. Often these decisions are dictated by constraints such as a job or appointment, but in many cases, it is left up to us. In these situations, social influence is a powerful motivator. We turn to our social contacts for trusted advice, access to new information, and because some choices are better if others make them as well. Choices related to mobility are no different. Where we choose to live and work can have large affects on our productivity and how we spend our time. Who we meet and talk in coffee shops and on busses can control our access to information. To understand how cities function, then, it is essential to understand how interactions between people drive their choices.

As with mobility, ubiquitous mobile computing is creating opportunities to study social interactions. Communication devices at their core, mobile phones now support hundreds of applications designed to make it easy to connect with others nearby and far away. Social interactions are captured in calls, emails, and tweets, while movement is logged by check-ins and GPS traces [166, 97, 190, 220]. Studied separately, social and mobility data have produced a wealth of insights. Our understanding of how information and diseases spread [221, 225, 235], how our friends affect our well being [48, 61], and how societies are structured [79, 234, 174, 81, 27, 22] has been

greatly improved by studying large social networks. Mobility data has revealed that human movement is regular, predictable [205, 202], and unique [71]. To complement empirical findings, a number of simple models have been proposed to reproduce the basic dynamics of both social networks [161, 32, 236] and mobility [202, 99, 200, 190], but the two have been traditionally treated as independent.

Recognizing the interaction between social behavior and mobility, researchers began measuring the relationship between the two. They found that social networks are heavily influenced by geography. We are far more likely to be friends with someone nearby than far away [136], a fact that is useful for predicting missing links [227, 66]. With an estimated 15% to 30% of all trips taken for social purposes, it is not surprising that the movement of our friends can improve predictions of where we will be next [58, 78, 99]. While insightful, the primary interest of most previous studies was measuring and reproducing patterns of geographic distance and its impact on network topologies [99]. In dense urban areas, however, distance is less restrictive. Residents have access to a variety of transportation options and are free to choose locations that provide the best goods and services rather than the closest. The self-organized districts and neighborhoods of cities make it more natural to describe mobility as movement between sets of locations, or habitats [13]. Which habitats users share with their contacts and when they share them may indicate the nature of the social relationship: e.g. a coworker or a friend [82]. Two individuals co-located between 9am and 5pm on weekdays likely have a different relationship than two who are found in the same area at midnight on a Saturday. In these scenarios, mobility is defined and measured as discrete visits to places within a city that are shared with different types of social contacts at different times and previous work has shown that users who visit similar places are more likely to be friends in online location based social networks [58].

Understanding these interactions these interactions is now having immediate and profound impacts on transportation. Car and ride sharing services like ZipCar and Uber are deploying algorithms to match individuals traveling to and from similar locations. As most trips can be shared and most cars sit ideal most of the time, the

mobility on-demand options have the potential to dramatically reduce the amount of capital and infrastructure required to serve travel demand. However, a better understanding of which individuals are good candidates to share is critical to scaling up these ideas.

Here we describe a set of metrics to explicitly measure patterns of mobility and social behavior that occur within the context of cities. Using call detail records (CDRs) produced by millions of mobile phone users, we find that individuals have far more similar visitation patterns to social contacts than to strangers and that the movement of these contacts can be used to reconstruct a considerable portion of the individuals' movements. We also find strong correlations between tie strength and mobility similarity and show that mobility similarity can be used to classify social relationships and recover semantic information about the nature of a link in the social network. Finally, we propose an extension to the mobility model described in [202] that incorporates movement based on the visitation patterns of social contacts and can reproduce empirical relationships found in the data. We call this model the GeoSim model and compare it against empirical data and two other mobility models. The generality of these results is demonstrated by their reproducibility in three different cities in two different countries. This study presents advances in the understanding of how social behavior affects our spatial choices in the context of information and communication technologies (ICTs).

3.2 Materials and Methods

3.2.1 Data

Call detail records (CDRs) are generated when a mobile phone user performs an action that requires the provider's network, for example placing a call or sending a text message. These records generally contain the ID of the tower the phone connected through, which gives a rough estimate of the user's location. When the individual receiving a call or message is a customer of the same provider, the unique identifier of

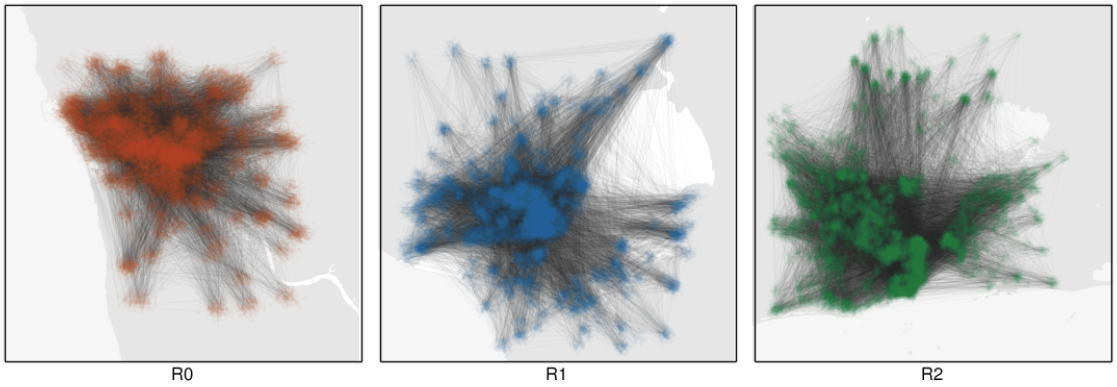

Figure 3-1: A small sample of calls between residents is shown for each of three cities. CDRs provide the location of each caller as well as record of communication between then. A dot is drawn at the approximate location of a user and a link appears between two users calling each other. Our aim is to identify useful and reproducible patterns from this coupled tangle of social and spatial behavior.

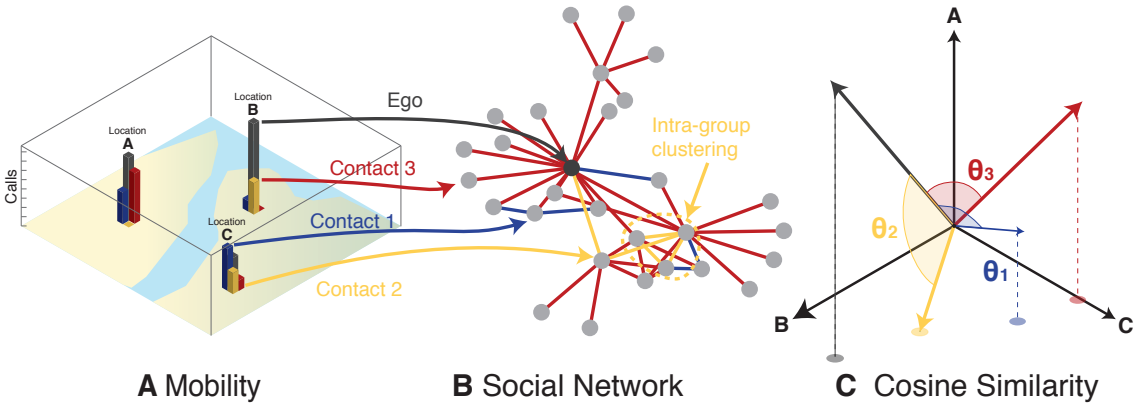

Figure 3-2: Similarity of visitation patterns between nodes in social networks. For each user, we keep track of (A) how many visits are made to locations across the city and (B) construct a social network by tracking calls to others. We can then define (C) the geographic cosine similarity between two users by computing the cosine of the angle between any two vectors in the location space.

the receiver and their location may also be stored. CDRs allow us to observe mobility patterns of individuals and construct social networks containing millions of people. Figure 3-1 shows a small sample of calls between city residents during a single hour and illustrates dynamics of the urban system we wish to understand.

Our data consist of anonymized CDRs collected from three cities (R1, R2, and R3) in two different industrialized countries. Two cities (R1 and R2) were obtained from the same provider in country 1, while another provider was used for the third

city (R3). The observation period covers 15 months in R1 and R2 and 5 months in R3 and contains over one billion events in total. Each record provides the time of the communication event, an anonymous unique ID for each caller and callee, and the ID of the tower used by at least the caller (in the case of R3) and in some cases the callee (R1 and R2). More information on the datasets can be found in Appendix 3.6.

3.2.2 Social and Mobility Measurements

In each city, we construct a social network containing all users (nodes) with sufficient call volume and connect users (edges) if they have regular contact between each other (see Appendix 3.6 for more detail). Each node is assigned a $48 \times L$ location matrix $\mathbf{L}$, where L is the number of unique cell towers in the city. Each row of this matrix corresponds to an hour of a typical weekday and hour of a typical weekend day (giving 48 hours in total) and each element $L_{t,j}$ contains the number of times that a user made a call from location j during hour t across the entire observation period (Figure 3-2A). We refer to individual rows of this matrix $\mathbf{v}(\mathbf{t})$ as *location vectors*. The location matrix and location vectors can be used to compute various mobility properties of nodes (mobile phone users). Summing all elements of the location matrix gives the number of calls made and received by a user $N = \sum_{t,j} L_{t,j}$ while summing each column and dividing by N provides the frequency of visits a user made to every location in the city, $f_j = \frac{1}{N} \sum_t L_{t,j}$. Summing visits to each location at all times gives a single location vector $\mathbf{v}$ for each user and represents the total visits made to each location over the period of data collection. Applying the sign function and summing across all elements of this vector provides the number of unique locations visited $S = \sum_j sign(v_j)$. All of these features are measures of a user's mobility behavior within the city.

We can also compare the location matrices and vectors of two mobile phone users and measure similarities between the two. While a number of metrics could be used to measure mobility similarity between nodes (Figure 3-2B), here we focus on the cosine similarity between the location vectors of two nodes i and j defined as: $\cos\theta_{i,j} =$

$\frac{\mathbf{v_i} \cdot \mathbf{v_j}}{|\mathbf{v_i}||\mathbf{v_j}|}$. The cosine similarity measures the cosine of the angle between two vectors in our L-dimensional *location space* (Figure 3-2C). It has been shown to correlate strongly with the probability of being friends in an online social network [58] and has a number of desirable properties. It is sensitive to visit frequencies rather than set intersections alone, so two users who share frequently visited locations appear more similar than those who share less important destinations. Unlike the Pearson correlation coefficient, it does not overstate similarity when vectors contain many zero elements (as is often the case) and finally, the cosine similarity is a measure of the angle only and is not affected by differences in the total number of calls made by two users. For the remainder of this chapter, we refer to the cosine similarity between two locations vectors as *mobility similarity*.

The mobility similarity between two users can be computed from their entire movement history or visits during a small portion of a weekday or weekend. In the former case, we assign a single mobility similarity value to an edge in the network, while in the latter, we assign a time series of cosine similarity $\cos\theta(t) = \frac{\mathbf{v_i(t)} \cdot \mathbf{v_j(t)}}{|\mathbf{v_i(t)}||\mathbf{v_j(t)}|}$. This time series reveals how often two users visit the same places at a given time of the day and will later function as an attribute to differentiate between types of social contacts.

Within this mathematical framework, we can calculate an upper bound on how much of an individual's location vector can be reconstructed from a linear combination of the location vectors of other users. For example, a co-worker may share office space with an individual, but not live in the same neighborhood, while the opposite may be true for a member of that individual's family. By combining the visitation patterns of the co-worker and family members, however, a complete picture of an individual's visitation patterns can be obtained. Mathematically, we define a set of users F for each individual i in the network. For example, we may choose F to be individuals that node i shared reciprocated calls with or a random set of nodes. The location vectors $\mathbf{v_j}$ where $j \in F$ are used as columns of an $|F| \times L$ matrix we denote as $\mathbf{A}$ and span a subspace of the L-dimensional location space. We then use QR-decomposition to find an orthonormal basis $B = q_1, \ldots, q_{|F|}$ for $\mathbf{A}$. Our target user's location vector is then

projected into this vector subspace: $\hat{\mathbf{v}} = \sum_{i=1}^{|F|} \langle q_i, \mathbf{v} \rangle q_i$. This projection represents the best approximation of a user's visits based on the visits of users in F. We can quantify how it compares to a user's true visitation patterns by taking the ratio of its magnitude with the magnitude of the actual location vector $|\mathbf{v}|$. We refer to this ratio as *predictability* and define it mathematically as $\frac{|\hat{\mathbf{v}}|}{|\mathbf{v}|}$. When predictability is 1, the visitation frequencies of a user can be completely obtained from location vectors of users in F and when it is 0, nothing about their visits can be learned. We note that for values between 0 and 1, predictability cannot be interpreted as the fraction of a user's visits that can be recovered as the vector norms are computed using the standard L2 norm. In principal, however, these two quantities should be strongly correlated because the individual elements location vectors can never be negative.

We next apply these methods and metrics to social network and mobility data from three cities.

3.3 Results

3.3.1 Correlations between social behavior and mobility

Though similarity can be measured between any two arbitrary nodes and predictability from an arbitrary set of nodes F, we hypothesize that an individual will likely be more similar to and predictable by social contacts. To test this, we compare the mobility similarity between users that call each other regularly with the similarity between random users and the predictability achieved using a node's social ties with the predictability using random sets of nodes (essentially rewiring the social network, but leaving mobility intact). Figures 3-3A and 3-3B show the distribution of similarity and predictability values for the networks in each city. We find significantly more similarity and predictability in empirical networks when compared to random re-wirings. The similarity distribution is bimodal, with peaks at very low similarity near 0 and very high similarity near 1. We measure very high values of predictability when using an individual's social contacts as opposed to a random set of people in

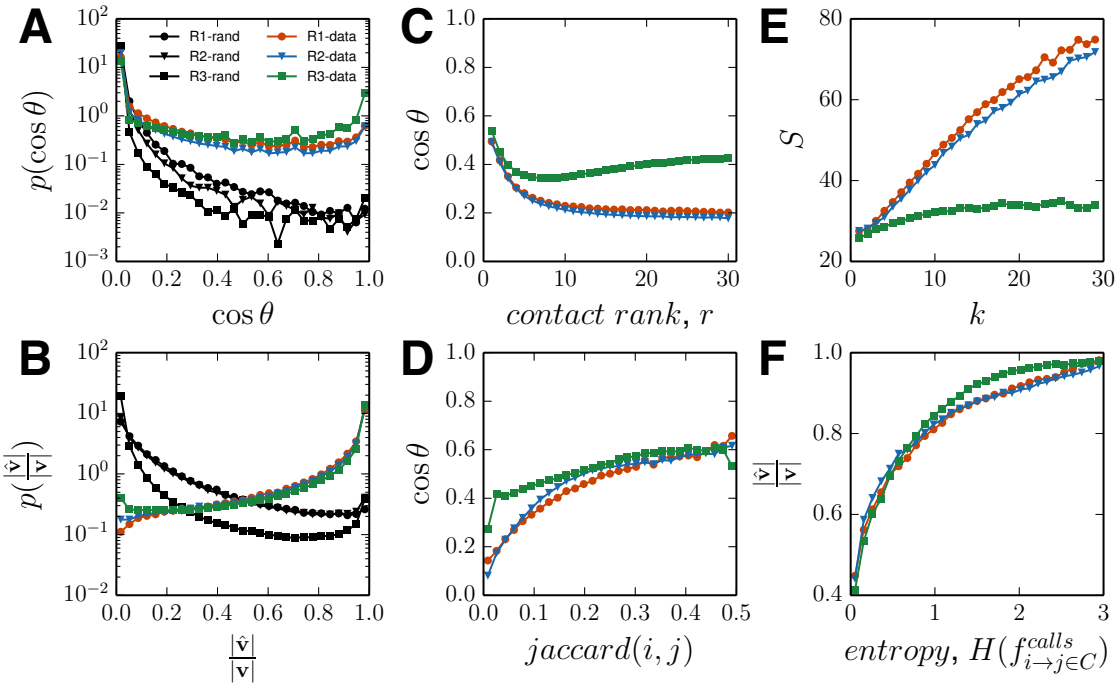

Figure 3-3: Correlations between mobility and social behavior. For each city, we compute the (A) distribution of cosine similarity and (B) predictability using observed edges (colored lines) and compare to distributions made using randomized edges. We find both mobility similarity and predictability are much higher when using actual social contacts compared to random users. Social similarity is also correlated with mobility similarity. (C) Ranking each user's contacts by number of calls, we find that stronger ties are more geographically similar. (D) Moreover, the more common contacts shared by two users, the more geographically similar those individuals tend to be. Finally, we explore how social behavior is correlated with mobility. (E) We find that users with more unique contacts tend to visit more unique locations. (F) Users who distribute their calls to contacts more evenly (higher entropy) are more predictable than users with more uneven call distributions. This suggests that users who share social attention more evenly also share locations. Figure S2 and S3 in the Appendix 3.6 show these results controlling for call frequency.

the same city. As other studies have suggested, we find that visitation patterns are strongly linked to our social relationships; our movements are far more similar to our social contacts than random users.

Interestingly, we observe higher levels of mobility similarity between users separated by short network distances. We find that two connected nodes are on average 10 times more geographically similar that two randomly selected nodes. Nodes separated by two hopes, or "friends of friends", are nearly twice as similar as randomly selected nodes and this elevated similarity is observed up to three hops from an individual

(see Appendix 3.6 Figure S5 for details). This result is expected as two users who do not contact each other may both visit the same friend.

Next, we explore the relationship between tie strength and mobility similarity. We rank all contacts that a user calls by the number of calls shared between them (1 being contact that shares the most calls) and compute the average mobility similarity for all edges with a given rank (Figure 3-3C). Stronger contacts have higher mobility similarity on average than weaker ties, though this effect subsides for contacts below rank 10. We note that region R3 shows a slightly different trend. This is likely due to the shorter observation period in this region resulting in few individuals with more than 10 regular contacts, biasing the tail of this distribution (see Appendix 3.6 for more details). We also observe a positive correlation between social similarity as measured by the Jaccard index between the neighbors of two nodes and mobility similarity (Figure 3-3D); individuals who share more social contacts share more locations.

We also find other aspects of social behavior to be correlated with mobility. Individuals with more friends tend to visit more locations, but despite this exploratory behavior, are still more predictable due to increased information provided by additional contacts to reconstruct these movements from (Figure 3-3E). Again R3 appears as an outlier due to the shorter observation period and the absence of mobility information on the user receiving a call. We then measure the entropy of the distribution of frequencies that a user i calls another contact j and find that individuals with more entropic calling patterns (distribute their calls more evenly) also visit more distinct places and are more predictable (Figure 3-3F). The visitations patterns of those who spread social attention more evenly can be more easily reproduced. Finally, to ensure that these results are not an artifact of sampling frequencies, we compute these distributions and correlations controlling for the number of CDR events by and the degree of a user, finding no change in the relationships (see Figures 3-7, 3-8, and 3-9 in the Appendix 3.6).

3.3.2 Contextualizing social contacts with mobility

Having demonstrated that social behavior and location choices are strongly correlated, we next use temporal variations in mobility similarity to provide context into the type of social relationship between two individuals in our networks. We measure mobility similarity $\cos\theta(t)$ over the course of a typical weekday and weekend under the hypothesis that different types of social contacts will have different levels of similarity at different times. To identify any groups, we use a simple k-means unsupervised clustering algorithm on these similarity time series. We find three persistent groups. While we have no ground truth data about the nature of these relationships, for clarity, we label each group according to its qualitative signature: (*i*) *acquaintances* with uniformly low levels of similarity, (*ii*) *co-workers* with high similarity during work hours on weekdays and low similarity on nights and weekends, and (*iii*) *family/friends* with high similarity on nights and weekends. Figure 3-4A shows the cluster centers for each group. While other interesting clusters are found for $k > 3$, they appear as subgroups of the three general archetypes we discuss here. More information on the clustering method along with results for different numbers of clusters and different clustering methods can be found in the Appendix 3.6. These three groups appear in each city despite the unsupervised nature of the algorithm; cluster centers start at random locations, yet find remarkably similar final positions in each city.

Assigning each edge to a cluster based on the time series of mobility similarity effectively paints all edges in the network a specific color as illustrated above in Figure 3-2B. Previous work has found that edges in real social networks are much more likely to be arranged in triangles, resulting in high clustering coefficients. In this case, we expect that some social groups, such as co-workers or close friends, should exhibit high degrees of intra-group clustering, while others such as acquaintances do not. For example, many of an individual's *co-workers* visit similar places during work hours and tend to call each other because they are part of the same office community. We find evidence of this when measuring the clustering coefficient within subgraphs containing only edges belonging to a single mobility similarity cluster (Figure 3-4B). Interestingly,

the clustering coefficient (C_g) of *acquaintances* is much lower than the *co-workers* and *family* ties despite consisting of nearly 70% of links in the network. This provides additional evidence that we are capturing very different types of relationships with our classifications based on mobility similarity. Moreover, these results highlight mobility similarity as a property to label functional communities within social networks as well as individual edges.

Next, we consider how the composition of an individual's ego-network correlates with their mobility. Is a person with a stable job and family is likely to be less exploratory and more predictable than a young college student with many acquaintances? To answer this, we bin nodes into groups based on two mobility metrics, the number of unique locations visited S and how predictable that user is $\frac{|\hat{\mathbf{v}}|}{|\mathbf{v}|}$. We then compute the fraction of edges that belong to each classification for all nodes in each mobility bin. Figure 3-4C shows that users who tend to visit more unique locations tend to have a higher fractions of *acquaintances* in their ego network, while Figure 3-4D suggests that less predictable individuals tend to have fewer contacts in this category. Conversely, less spatially explorative individuals and individuals that are easier to predict tend to have higher fraction of *co-workers* and *family/friends* labels in their ego network. These results again show the ability of mobility similarity to add contextual attributes to a network and reveal novel relationships between the structure of a user's ego network and their mobility behavior. In future works, it may be interesting to explore correlations between the mix of one's ego network and social behaviors such as their propensity to form new contacts [159].

3.3.3 Coupling social ties and mobility

Given the clear empirical relationship between social contacts and mobility, our remaining task is to identify a coupled model that captures these dynamics. While a number of models consider mobility alone [202, 200, 97], only a few have attempted to link the two [99, 58]. Those that have combined social and mobility behaviors have consistently found nearly 15-30% of trips are made for social purposes. Though these coupled modeled have had considerable success reproducing patterns of geographic

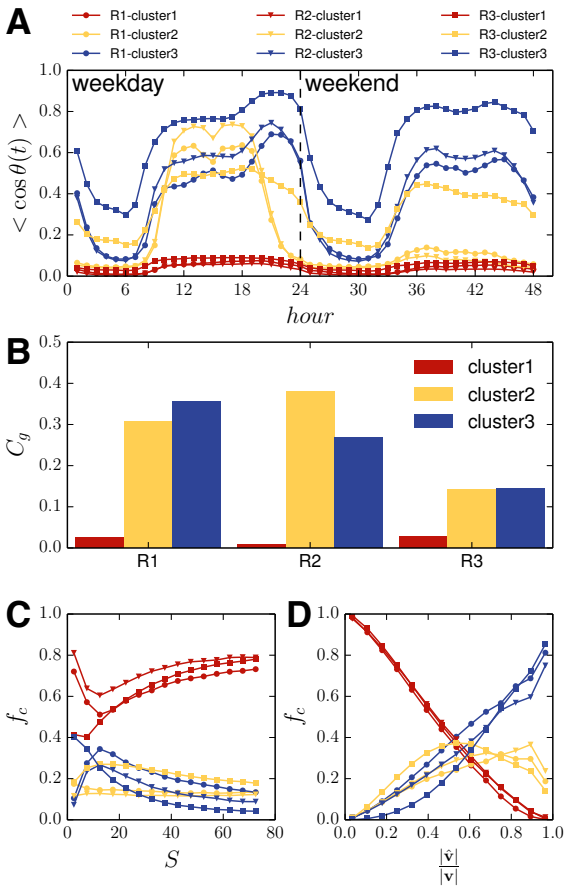

Figure 3-4: Characterizing social ties based on similarity of movement over time. (A) We perform k-means clustering on the set of similarity time series from edges in the network. We find three groups emerge in each city: (*i*) *acquaintances* who have low levels of similarity across all times, (*ii*) *co-workers* who have elevated similarity during work hours on weekdays, but lower levels on weekends, and (*iii*) *family/friends* who have high similarity on nights and weekends. (B) For each city we construct subgraphs containing only edges in a single cluster. We find that these subgraphs retain high clustering coefficient (C_g) within the co-worker and family/friend group while acquaintances are far less likely to have ties among each other. Finally, we explore how an user's behavior correlates with the mobility characteristics of their immediate social network. (C-D) We group nodes based on their mobility characteristics (unique locations visited S and predictability $\frac{|\hat{\mathbf{v}}|}{|\mathbf{v}|}$) then compute the fraction of edges that belong to each of the identified clusters for each node in the group. Individuals that are more exploratory (visit more unique places) tend to have higher fractions of *acquaintances* ties than individuals with lower mobility while the reverse trend is observed for the most predictable individuals.

distance within social network structure, but, as we show, do not always capture properties of geographic similarity.

In light of the time scales we are studying, we make the assumption that our social network is static and extend the mobility model introduced by Song et al. [202] to

include movement choices based on social contacts. We call our extension the *GeoSim* model[1]. We compare our model to the original individual-mobility model (IM model) by Song et al. and the Travel-Friendship model (TF model) described by Grabowicz et al. [99]. See the Appendix for more details on implementation and parameters for model comparisons.

The GeoSim model works as follows: first, a population of N agents are initialized and connected to replicate the undirected social network constructed from the CDR data in R1. Each edge that exists in the call data, exists in the model, but all weights and similarities are set to 0. Agents are randomly assigned to a location at the start and their location vectors are initialized to reflect this single visit. They are allowed to move in a discrete space of L locations replicating the towers from CDRs.

Each time step corresponds to a single hour of the day. At each time step, individuals decide whether or not to change locations according the waiting time distribution measured in [202], a power-law with an exponential cutoff $p(\Delta t) = \Delta t^{-1-\beta} \exp(\Delta t/\tau)$ where $\beta = 0.8$ and $\tau = 17$ hours. If an individual moves, they must decide to either return to a previously visited location with probability $1 - \rho S^{\gamma}$ or explore and visit a new one with probability ρS^{γ}, where S is the number of unique locations they have visited thus far and $\rho = 0.6$ and $\gamma = 0.6$ are parameters chosen by procedures outlined in[202]. In the original model, an individual u *preferentially returns* to a location l with probability proportional to the frequency of previous visits, $P(l) \propto f_l^u$ and new locations to explore are chosen uniformly at random (note that in our version of the model distance is irrelevant).

In our extension of this model, we choose some locations based on social influence. When picking a return location, our agent has two possibilities. With probability $1-\alpha$, they select a return location with the preference for locations they have visited in the past as in the original model. With probability α a social contact v is chosen. The probability a given contact is chosen is directly proportional to the current mobility similarity between the two, $P(v) \propto \cos(\theta_{u,v})$ and a location to visit is chosen based on

[1]We have released code and data required to run this model online at http://humnetlab.mit.edu/wordpress/downloads.

a preference to visit locations frequented by the selected contact, $P(l) \propto f_l^v$ (note the location choice is repeated until an agent finds a location they have visited before). In the social case, this amounts to preferential return based on a contact's visit frequency as opposed to the ego's visits. In the event that an agent is exploring a new location, the same weighted social coin is flipped. This time, though, with probability $1 - \alpha$ a random, previously unvisited location is selected and with probability α the agent again chooses a contact based on mobility similarity and chooses a new place to visit based on the visit frequencies of that contact. The cosine similarity across all edges is computed and updated over as the model progresses and changes dynamically during the simulation. A schematic of this process can be found in Figure 3-5.

In this variant of the mobility model, the parameter α controls the influence of social contacts on the visitation patterns of individuals. When $\alpha = 0$, we recover the original mobility model of [202], while when $\alpha = 1$ all location choices are influenced by social ties. In reality, each user may have an inherent value of α that we cannot observe. To incorporate this heterogeneity, we simulate this model for a number of distributions of the parameter α. We find an exponentially distributed α with a mean of $\langle\alpha\rangle = 0.2$ produces a close fit to distributions of mobility similarity and predictability observed in the population and refer the reader to the Append for results for different distributions of α. This value is consistent with the results of both Cho et al. [58] and Grabowicz et al. [99] who find that roughly 15-30% of trips were motivated by social intentions.

Having found an appropriate distribution for α, we next compare simulation results with this distribution to results from the IM model (equivalent to the GeoSim model with $\alpha = 0$) and the TF model all run for the same 1 year duration and populations size. Like the IM model it extends, the GeoSim model is able to reproduce elements of individual mobility such as the rate of exploration of new locations $S(t)$ over time (Figure 3-6A) as well as frequency at which users visit their locations f_k (Figure 3-6B). Here the TF model adequately reproduces exploration rates, but produces a flatter visit frequency distribution. In the case of mobility similarity and predictability, however, only the GeoSim model reproduces observed behavior (Figure

3-6C-D). Interestingly, the TF model results in relatively high predictability of users, despite similarity values orders of magnitude lower than those observed in the data or with the IM model. This is likely due to the flattened frequency distribution which the cosine similarity is highly sensitive to. Even if two users share a few locations due the friendship component of the TF model, there are preferential dynamics that will continually bring those two users back to that place, increasing cosine similarity. On the other hand, this flat frequency distribution makes it highly likely that users will share at least some locations in commons with each other, making it possible to reproduce location vectors based on social contacts. Despite its inability to recover these distributions, the TF model is the only model tested that builds a social network endogenously. For this reason, we hope future work will find variants on this model capable of dynamically reproducing empirical data of both social and mobility behavior.

3.4 Discussion

Linking mobility to social ties has generated a number of insights into the dynamics of both. Social networks are embedded in geography where face-to-face interactions are often preferred and chance of interacting with those nearby is greatest. At the same time, we are willing to travel to achieve this proximity and rendezvous at places across the city for work and play. Novel high resolution data sets passively collected from mobile, online devices now enable us to quantify the relationship between mobility similarity and social behavior. Here we have offered new metrics and empirical findings that relate social behaviors to mobility similarity and predictability. Our results show that our mobility is far more similar to our social contacts than strangers and that this similarity can be used to reconstruct our own mobility patterns. We find strong, positive correlations between tie strength and mobility similarity. Moreover, temporal variations in this similarity reveal three distinct groups of social ties that hint at semantic types of relationships such as co-worker or family member. These subgraphs often have high levels of intra-group clustering, suggesting func-

tional groups of individuals within the network. The mix of these groups amongst the edges of an individual's ego network is correlated with their mobility behavior; users with many dissimilar contacts tend to explore more locations. Speaking to their generalizability, these results persist across three different cities in two countries.

Finally, we extended an established mobility model to include choices based on social behavior that replicates the empirical findings described here as well as from other works. We call this model the GeoSim model and have compared its results to two similar models. We hope that this model provides a useful tool for future work in the area. The findings presented have a number of implications for those interested in social networks or mobility applications extracted from ICTs. Additional contextual information of relationships may help predict missing links or provide critical details to more accurately model of the flows of information or diseases. Urban planners or those needing good estimates of travel demand can incorporate social mechanisms like the ones described here to improve on their models and to capture movements previously unaccounted for. Robust findings that classify social contacts from passive data alone may influence future studies and help with data informed policies through city science. In the new data rich reality of cities, deeper insight into the connections between us will help make the places we live more sustainable, efficient, productive, and fun.

3.5 Acknowledgments

The work in this chapter was the result of a collaboration with Carlos Herrera-Yagüe, Christian M. Schneider, and Marta C. González. I thank them for their advice and support.This work was partially funded by the BMW, the Accenture-MIT alliance and the Center for Complex Engineering Systems (CCES) at KACST under the co-direction of Anas Alfaris. Jameson L. Toole would like to acknowledge funding awarded by a National Science Foundation Graduate Research Fellowship.

3.6 Chapter 3 - Appendix

3.6.1 Data

Our data consist of anonymized call detail records collected from three cities (R1,R2, and R3) in two different industrialized countries. The same provider was used for the two cities in the same country (R1 and R2), while another provided the remaining city (R3). In R1 and R2, data cover 15 months while R3 contains 5 months of data. In total there are over 1 billion events contain the time and duration of a communication event between a caller and callee as well as the towers used by one (in the case of R3) or both (in the case of R1 and R2) of the users. Though data sharing agreements to protect privacy prevent us from sharing the locations of each region, they are major metropolitan areas with densities closely matching that of Boston, MA, USA.

3.6.2 Social Network Extraction

To build the social networks for each city, we employ the following procedure. First, we consider only users that appear in over 200 communication events within each city's metro region over the course of the entire data collection period. Second, we only draw an edge between two users if they make more than two calls between them during that time. Properties of the three networks as well as the number locations (cell towers) within each metro region can be found in Table 3.1.

Table 3.1: Basic statistics on the networks and spatial extent of each region considered.

City	Nodes	Edges	$\langle k \rangle$	Towers
R1	133,587	997,287	14.9	249
R2	183,486	2,487,661	27.1	447
R3	635,731	4,197,093	13.2	935

3.6.3 Metric Definitions

Cosine Similarity, $\cos\theta$ In this work the cosine similarity is defined as the cosine of the angle between the location vectors of two users, $\cos\theta_{i,j} = \frac{\mathbf{v_i} \cdot \mathbf{v_j}}{|\mathbf{v_i}||\mathbf{v_j}|}$. In general,

cosine similarity can take values from -1 to 1, but in our case it is non-negative as it is impossible for location vectors to have negative elements. This restricts angles to between 0 and $\frac{\pi}{2}$. We use the cosine similarity as a measure of how similar visitation patterns are between two users.

predictability, $\frac{|\hat{\mathbf{v}}|}{|\mathbf{v}|}$ Predictability provides an upper bound on how much of a target user i's visitation patterns can be reconstructed by the visitation patterns of a set of other users, F. In general, F can be made up of any users, but here we define it as the set of social contacts called by user i. Location vectors exist in an L-dimensional location space, where L is the number of unique locations that can be visited. In practice, most users visit only a small number of possible locations and the locations vectors of the users in F typically span only a subspace of the entire location space. If this subspace contains most of user i's location vector, then we can reconstruct their mobility patterns with a high degree of accuracy. If, however, user i's location vector is orthogonal to each vector in a basis of this subspace, then none of the user's visits can be recovered.

To quantify this, we first construct a $|F| \times L$ matrix $\mathbf{A} = [v_{f_1}, v_{f_2}, \ldots, v_{f_n}]^\intercal$ where f_j are contacts in F. We next use qr-decomposition to construct an orthonormal basis $B = q_1, \ldots, q_{|F|}$ of $\mathbf{A}$ that spans the subspace of the entire L-location space that is defined by the users in F. We can then project the original location vector of user i into this subspace to create the best approximation $\hat{\mathbf{v}}$ of that user's visitation patterns that can be constructed by users in F: $\hat{\mathbf{v}} = \sum_{j=1}^{|F|} \langle q_j, \mathbf{v} \rangle q_j$. To measure the accuracy of this approximation, we take the ratio of its magnitude to the magnitude of the original location vector for the target user, $\textit{predictability} = \frac{|\hat{\mathbf{v}}|}{|\mathbf{v}|}$. Predictability can take values between 0 and 1 where the former indications none of the user's visits can be reconstructed by linear combination of users in F and 1 indicates all of a user's visits can be recovered.

We caution, however, against interpreting intermediate values of predictability

as “fraction of visits correctly predicted". The magnitude of vectors are computed using the L2 norm are thus not equivalent to comparing percentages of visits recovered. The two values, however, are highly correlated in this case as there can be no negative elements in location vectors or their approximations. Finally, we note that predictability here is an upper bound approximations to a user's location vector using a linear combination of their contacts visits. Computing predictability requires full knowledge of a user's location vector. In the absence of this information, some proxy must be identified to replace the coefficient $\langle q_j, \mathbf{v} \rangle$. We encourage future work exploring this.

Number of Unique Locations Visited, S The number of unique locations visited by a user is denoted as S. It can be computed directly from a user's location vector as $S = \sum_i sign(v_i)$.

Degree, k A user's degree in the social network is denoted as k.

Contact Rank, r For each user, we assign a contact rank r as a measure of tie strength to every neighbor in a user's ego network. We rank every contact of a user according to the number of calls made between them and assign a value of $r = 1$ to the contact exchanging the highest number of calls with the ego. We note that contact rank is not necessarily symmetric. User i may rank as user j's 3rd most called contact while user j is only user i's 10th most called contact.

Jaccard Similarity The Jaccard similarity is a measure of set intersection. It is computed as the fraction of elements from set A that exist in set B. In the context of our social network, it is a measure of tie strength is defined as the fraction of contacts shared by two users i and j: $jaccard(i,j) = \frac{F_i \cap F_j}{|F_i||F_j|}$, where F_i is the set of neighbors contacts of user i.

Entropy, $H(f^{calls}_{i \to j \in C})$ Entropy is traditionally used as a measure of randomness or disorder. In the context of this work, we measure the information entropy of various probability distributions. Distributions with probability mass spread more evenly across all possible outcomes are considered to be “more random"

and have higher entropy than distributions where a single outcome is far more likely than all others. Shannon's information entropy for a random variable X is computed as $H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$. In our case, we measure the entropy of the distribution of call frequency of a user to his or her social contacts. High values of entropy are calculated when a user distributes their calls evenly to all social contacts while low values are observed when they call a small number of contacts far more than the rest.

3.6.4 Controlling for number of calls

While mobile phones make excellent passive sensors of social behavior and mobility, they suffer from non-uniform sampling rates. Information is only recorded when a user uses the device leaving more observations at certain times of the day or week than others. Moreover, different users may use their devices more or less depending on habits or socio-economic variables. Because of this, we are careful to ensure that any metrics we measure in the data are not biased by different sampling rates.

Figure 3-7 shows the distributions of four metrics in each region for groups of users with similar numbers of calls over the observation period. In general, we find that calling frequency of users does not affect these distributions with the exception of the number of unique locations visited S, which increases with calling frequency. However, even in these cases the shape and trend of the distribution remains the same for each group with only the means shifted. Finally, we note that for region R3, the number of unique locations visited takes on a slightly different shape than regions R1 and R2 due to the fact that we only obtain location information for callers in this city and not for receivers as well. Our new metrics of mobility, cosine similarity and predictability, are least affected by different sampling rates.

We perform the same analysis for correlations between social behavior and mobility. For users with a given number of calls, we correlate their social metrics such as degree or the entropy with which they distribute calls to contacts with mobility metrics. We find that, similar to the distributions, most of these correlations do not depend on the number of calls made by a user. In cases where there is dependence,

the trends hold within groups of users that make the same number of calls (Figure 3-8).

3.6.5 Controlling for Degree

We measure the entropy of the distribution of calls that each user makes to his or her contacts. Users with higher entropy spread their calls evenly amongst social ties, while lower entropy means most calls go to fewer. The degree of a node sets an upperbound on the entropy a user can have. Thus, any correlations we measure may be biased by differences in the degrees of each user. To control for this, we plot correlations of call entropy to other metrics for groups of users with the same degree. Figure 3-9 shows that these trends are unaffected by differences in degree.

3.6.6 Social Distance and Geographic Similarity

We compute the average cosine similarity between two nodes separated by a social distance of k hops. Much like previous studies of homophily within social networks, we find that geographic similarity is elevated for two individuals who call each other, but this increase in similarity extends outward up to three hops away after which users are as similar as they would be to random users (Figure 3-10).

3.6.7 Clustering

The k-means clustering algorithm must be seeded with the number of clusters to find a-priori. In order to identify a reasonable number of clusters, we run the algorithm for multiple values of k and examine the resulting clusters as well as the silhouette score for each choice. The silhouette score decreases as the number of clusters increases indicating that there is little added benefit from additional splitting (Figure 3-11). Moreover, when examining the centroids of clustering results, the three main clusters identified break into similar groups that show small differences such as on weekends or in absolute similarity level (Figure 3-12). To ensure our results are not an artifact of the clustering algorithm chosen, we also perform clustering using a hierarchical,

agglomerative clustering technique using Ward linkage. In each region, we obtain clusters that match those found with k-means very closely (Figure 3-13).

3.6.8 Ego-network link type mixes

To ensure that our results are not an artifact of call frequency we plot the mix of an individual's social network as a function of the number of calls they make (Figure 3-14).

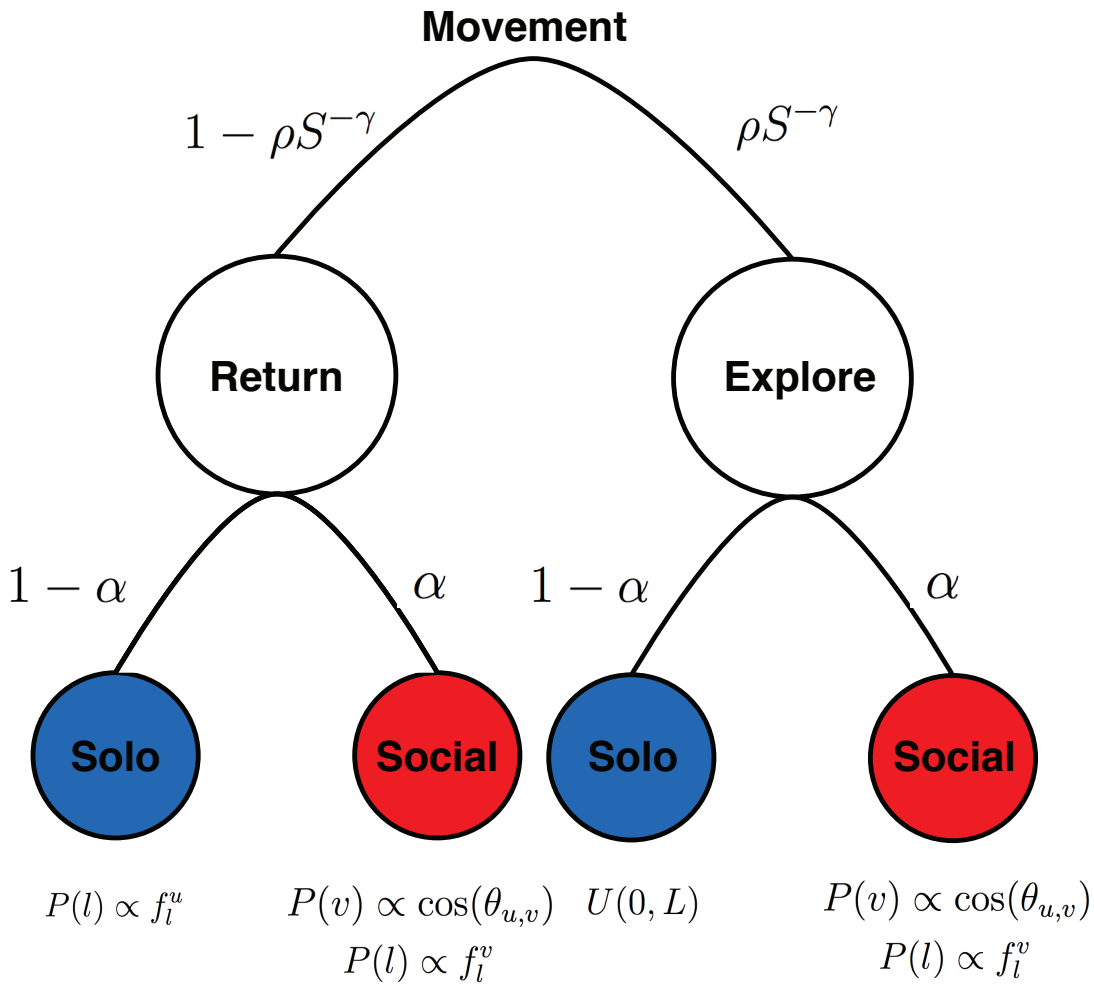

Figure 3-5: A schematic description of the GeoSim model. As in the IM model presented by Song et al., individuals first decide whether to return to a previously visited location or explore a new location. The actual choice of location to visit, new or returning, is made based on either a social influence with probability α or individual preference with probability $1-\alpha$.

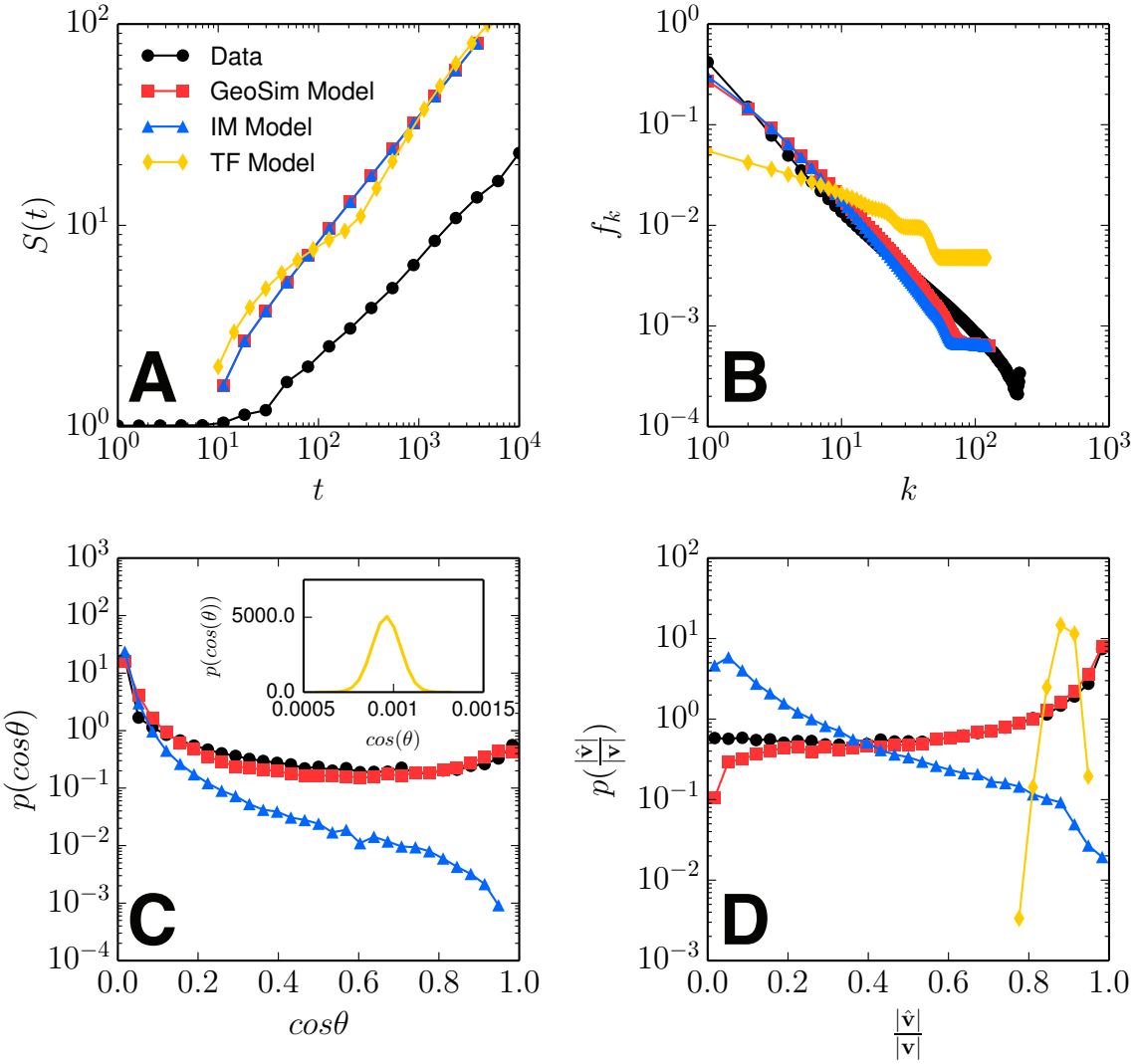

Figure 3-6: Comparing social mobility models. A) We compare model results simulating the rate of exploration $S(t)$ compared to empirical data. While all three models appear to estimate more absolute locations visited, the rate of this growth is consistent between them and in-line with data. B) For each user, we sort locations based on the number visits and compute the frequency that a user visits a location of rank k. We find that the IM models and our extension to it reproduce this distribution well, while the TF model is much flatter, distributing visits more evenly over all locations. C) Only the GeoSim model is able to reproduce patterns of mobility similarity and D) predictability. The TF model results shown in the inset in C shows similarity values orders of magnitude below the observed data. As the similarity is heavily influenced by the frequency distribution of visits, this deviation is likely due to the flatter distribution of f_k produced by the TF model.

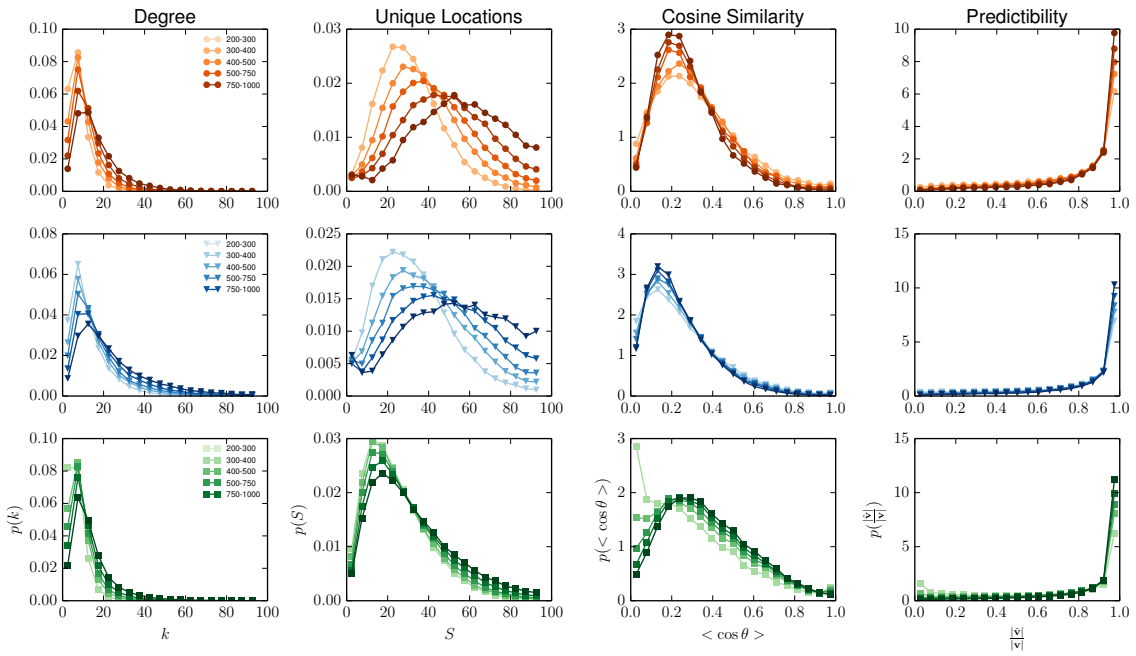

Figure 3-7: Distributions of different variables (columns) for each of the three regions (rows) for groups of users with different numbers of total calls. To ensure that measurements are not simply artifacts of differences in the amount a users interacts with their phone, we plot distributions of variables for groups of users with different activity levels. Users are binned first according to the number of records they have in the data set, then distributions of various mobility and social metrics are plotted for each user group. In general calling frequency does not affect these distributions with the exception of the number of unique locations visited where the mean is shifted right for users with more calls.

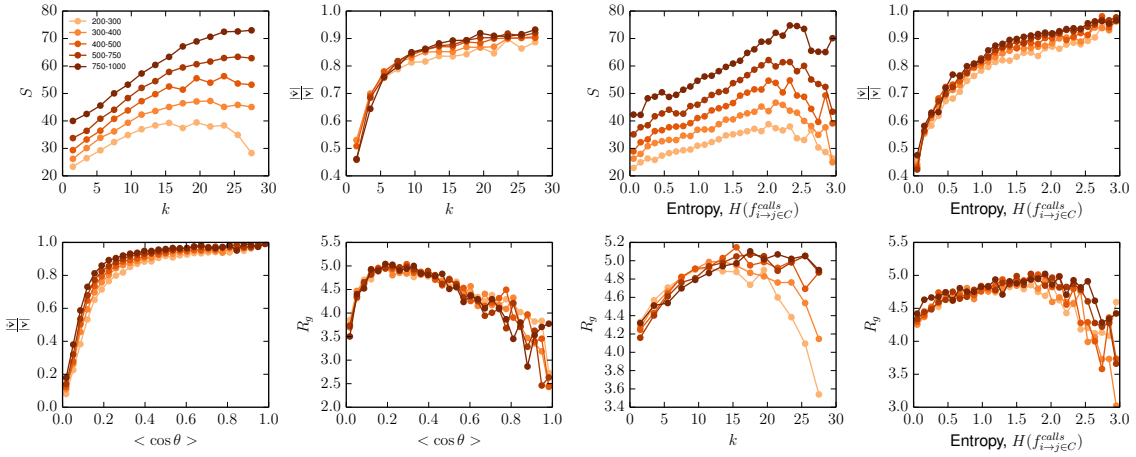

Figure 3-8: Various correlations metrics related to social behavior and mobility while controlling for the number of calls made by each user. Again, we bin users based on the number of records they have in our data set and then measure correlations between social and mobility metrics. We find, as was the case with distributions, these correlations are unaffected by sampling frequency.

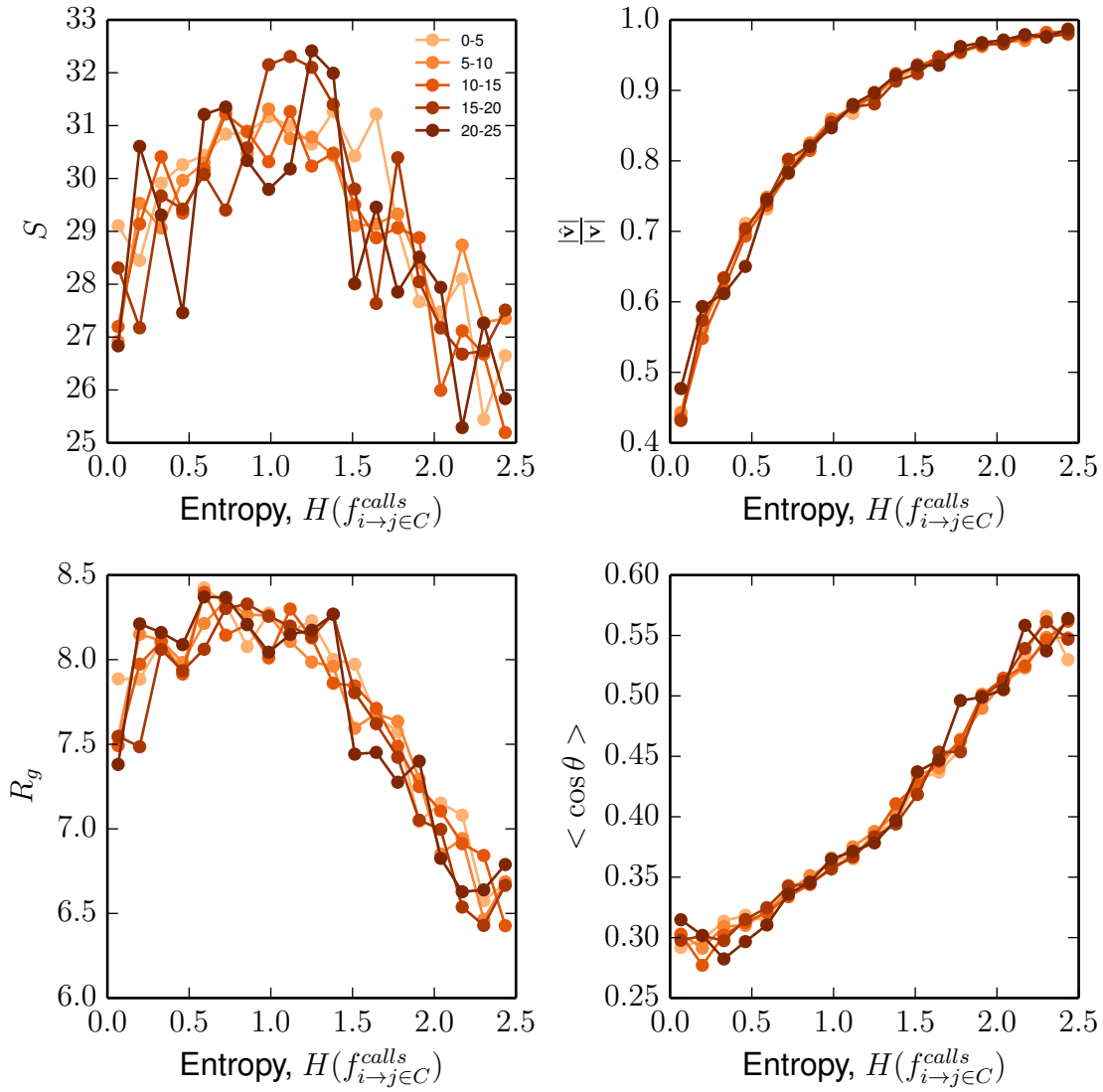

Figure 3-9: Correlation between the entropy of a node's call frequency distribution to contacts and mobility variables may be affected by the degree of each node. Measures such as entropy and predictability will naturally be affected by the number of contacts each user has. For example if a user contacts for people, the maximum entropy of the distribution of call frequencies to those individuals will naturally be higher than a user who has few friends. To ensure our correlations are not artifacts of the number of contacts each user has, we plot these correlations for groups of users with the same degree and show that these relations still hold.

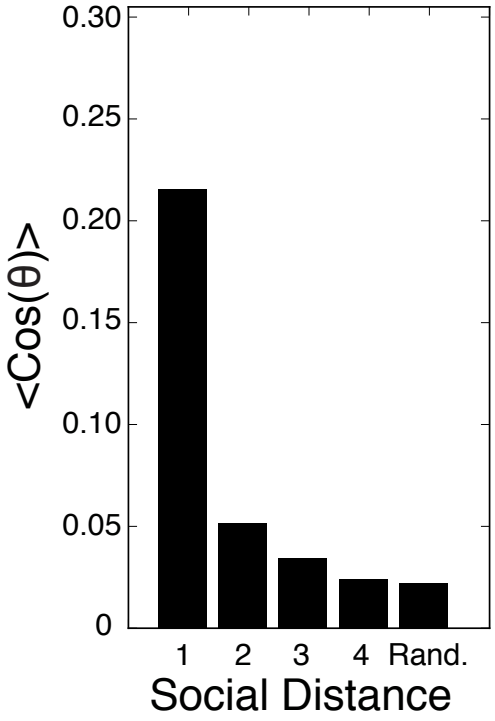

Figure 3-10: Social distance and geographic similarity. Nodes who contact each other are far more similar to each other than two randomly selected nodes. Here we compute the average mobility similarity between nodes separated by a certain number of hops. Even for nodes separated more two or three hopes, we elevated levels of similarity when compared to two randomly selected nodes in the network.

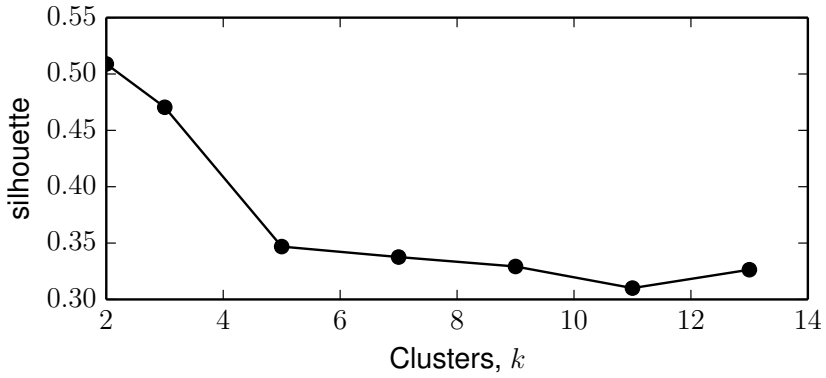

Figure 3-11: The silhouette score for different numbers of clusters. The silhouette score is a measure of the ratio between intra- and inter-cluster variance that gives a rough measure of the quality of clustering results (higher is better). The score drops steadily from the chosen number of clusters, 3, indicating that little is gained by additional splitting.

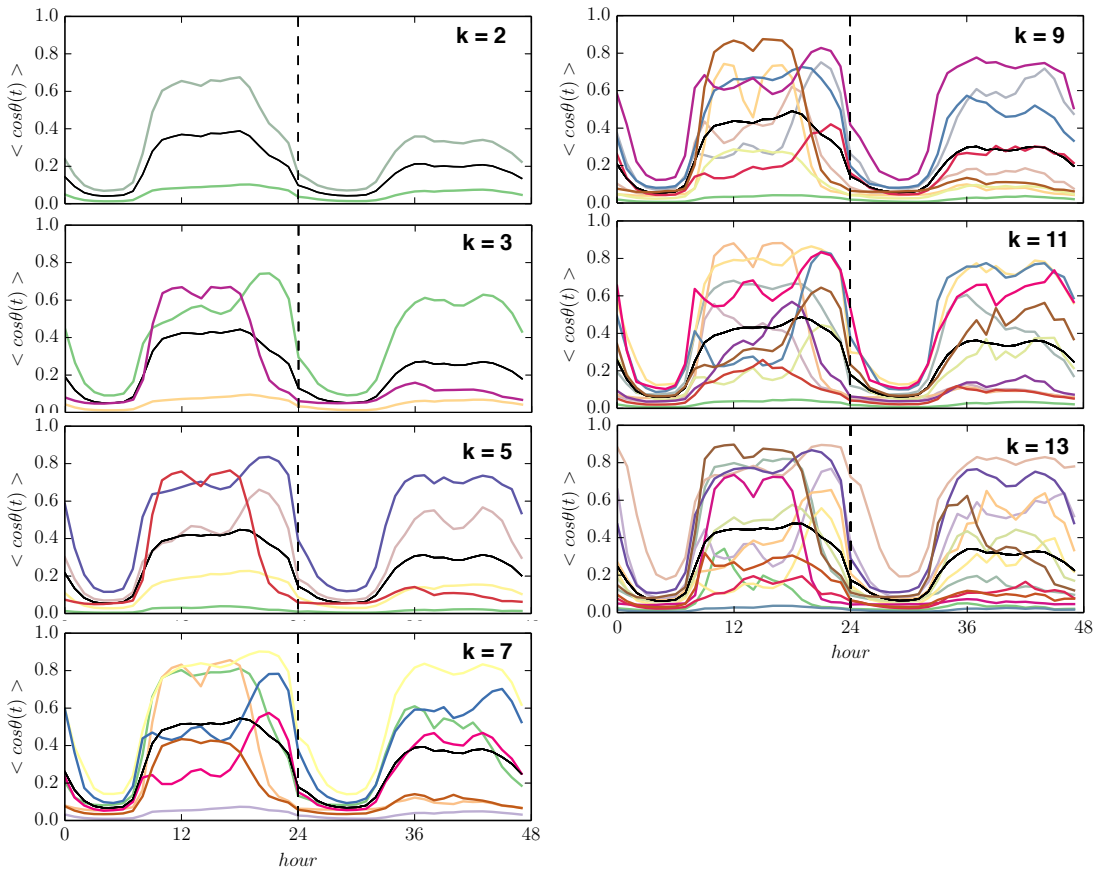

Figure 3-12: k-Means clustering results for various values of k in city R1. We perform k-means clustering for multiple values of k as a manual check that our choice of 3 clusters is appropriate. In general, additional clusters appear to be variations of three main themes used in the main text.

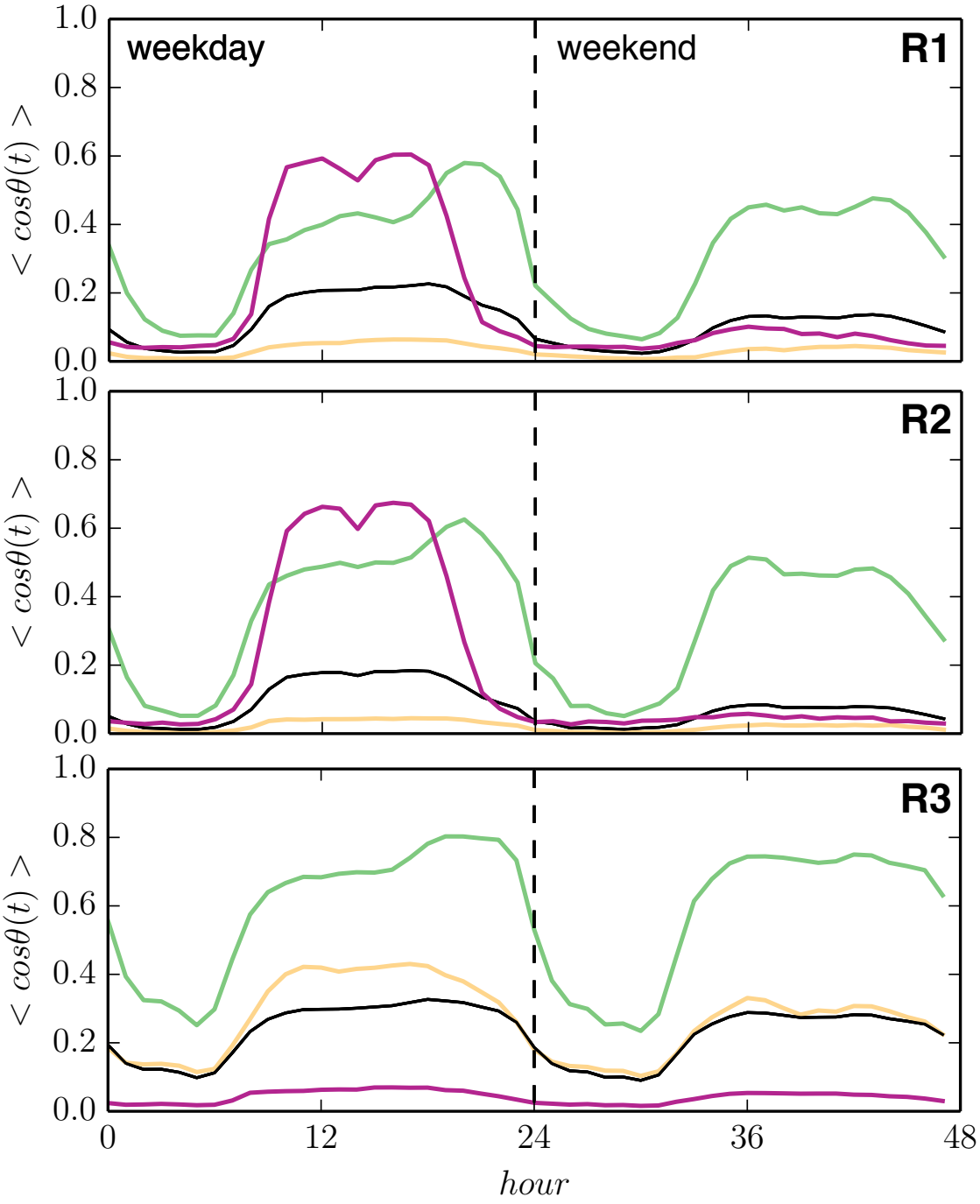

Figure 3-13: Results from a hierarchical, agglomerative clustering algorithm with Ward linkage. This clustering method clusters nodes based on connecting data points together if they are within some distance of one another and then examining connected components. The clusters in each region closely match results from k-means, suggesting that our results are robust to the exact clustering algorithm used.

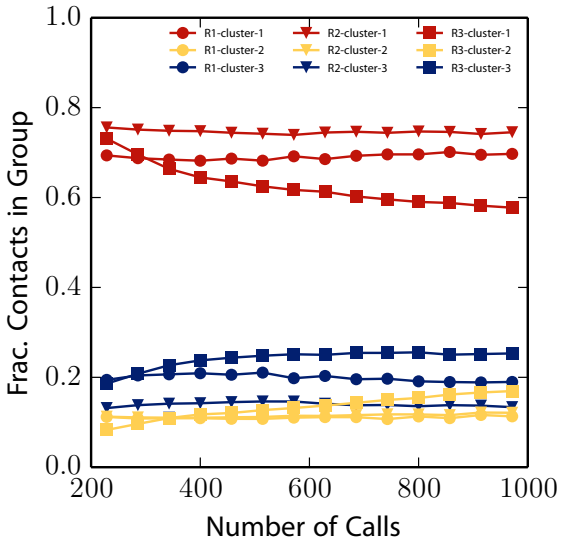

Figure 3-14: The mix of a user's ego network versus the number of calls they make. To ensure that the relationship between the mix of a user's ego network and their mobility isn't simply due to the fact that user's with different numbers of records having different edge mixes, we plot the average make-up for suers with different numbers of calls. We find that regardless of a user's calling frequency, the makeup of their social contacts is stable.

3.6.9 Model Comparisons

GeoSim Model

To account for heterogeneity in how social different person's are, we allow each individual in our simulation to have a value of α drawn from a distribution. We run the GeoSim model with various distributions of α to determine which best approximate measurements from the data. Figure 3-15B shows the various distributions of α we simulate while Figures 3-15C and D show resulting similarity and predictability distributions when compared to empirical measurements. We find that exponential distributions $p(\alpha) \propto \exp(-\lambda\alpha)$ result in distributions that match the general trend of the empirical distributions. We expect the precise parameter, λ, to vary from city to city or culture to culture, but values between 0.1 and 0.3 produce adequate results and are consistent with previous works that find roughly 15%-30% of trips are made for social purposes.

Individual Mobility Model (IM Model)

In their work, Song et al. [202] present a model for individual human mobility that we extend to the GeoSim model presented here. When $\alpha = 0$ for all individuals, we recover Song's model. The IM model relies on a 2 parameters, ρ and γ which control the propensity for a user to explore a new or return to a previously visited location. By varying these parameters, the authors can generate a range of mobility behaviors related to the rate of exploration and the frequency that users visit locations. We find that values of $\rho = 0.6$ and $\gamma = 0.6$ produce reasonable fits to both the exploration rates ($S(t)$) and the frequency that users return to locations f_k. We leave distributions such as the waiting time distribution measured by Song intact with $\beta = 0.8$.

Travel-Friendship Model (TF Model)

The model presented by Grabowicz et al. [99] proposes to model mobility and the growth of social networks simultaneously. In this model, geographic space is treated as a very small grid with cells δ on a side. At each time step, a user makes choices related to travel and friendship. For travel, an individual chooses to travel to the location of a random friend with probability p_v or with probability $1 - p_v$, chooses to jump to a new location. In the event of a jump, a distance is chosen based on the distribution measured by Song et al. and the individual then surveys all grid cells at that distance and chooses one to move to proportional to the population density of the cells. After a user has made a travel decision, they make choices related to friendship. For each other person within the individual's grid cell, a link between them is created with probability p, and with probability p_c a link is created with a random person anywhere. These dynamics reproduce social network distributions as well as distributions of distance between friends.

We implement the above model as described, but add one additional step to make it comparable to the CDR data used in this study and modeled by Song et al. The δ grid cells in the TF model are on the order of 100m $\times$ 100m in the original implementation. This is far smaller than the coverage areas of towers within a city

save for the very dense downtown areas. To make the the data from the two models directly comparable, we simply assign a users location to the tower that covers the grid cell they have jumped to. This preserves all the original behavior of the TF model, while making it possible to perform a fair comparison.

Finally, the TF model has four parameters: δ, p, p_v, and p_c. We set these parameters to the middle of the ranges estimated by Grabowicz et al: $\delta = 0.001$, $p = 0.1$, $p_v = 0.15$, and $p_c = 10^{-3}$. We run this model for the same population size and length as the IM model and GeoSim models.

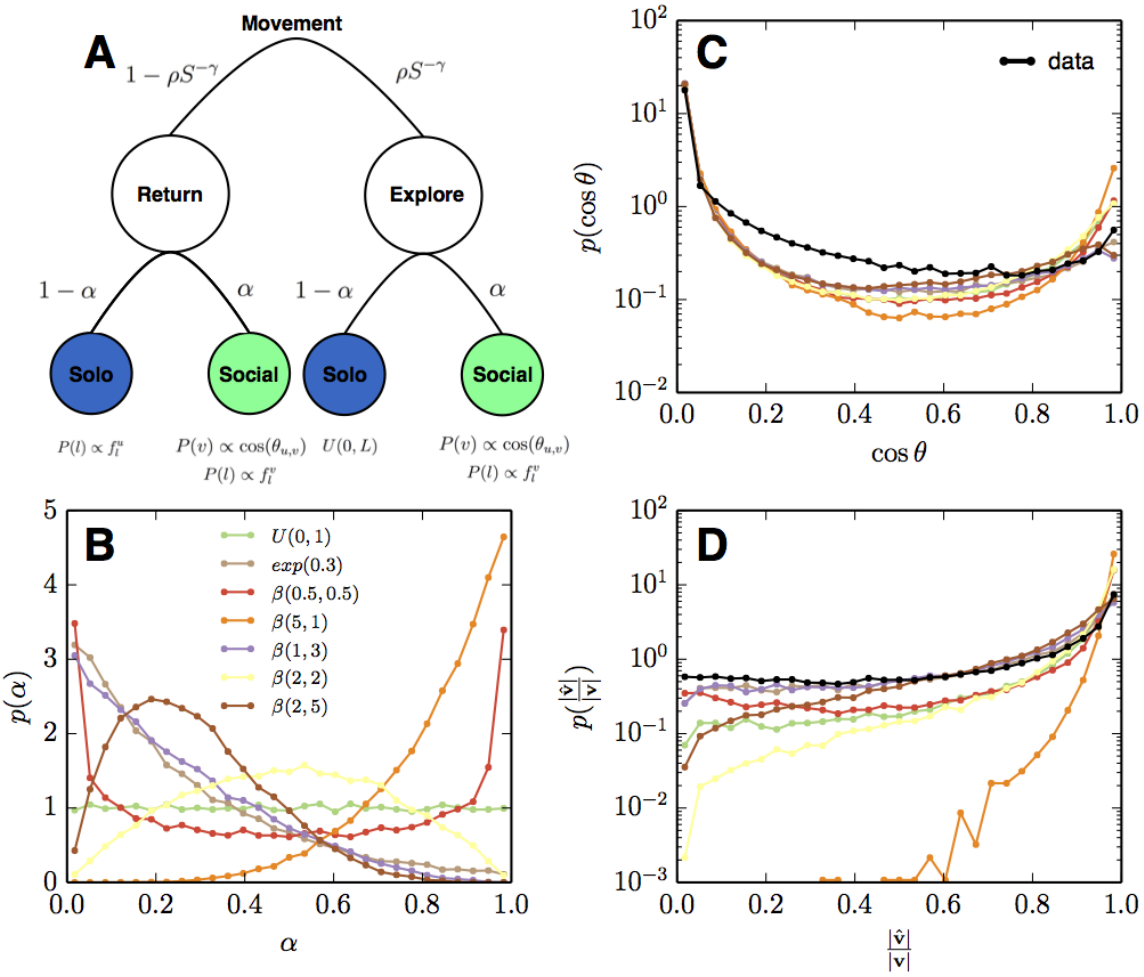

Figure 3-15: Our extended mobility model. (A) A diagram showing the choices made by an individual in deciding where to move next. We compare simulation results for different values of social influence α to distributions of (B) similarity and (C) predictability found in real data. The bimodal similarity distribution is recovered for higher values of α while the predictability results suggest that this parameters may vary for individual to individual resulting in a mix among the whole population.

Chapter 4

Economic Behavior in Cities

Cities are powerful economic engines. In the United States, more than 85% of the countries GDP is generated by the 78% of people living in urban areas with populations over 150,000[1]. At their founding, the efficiency of water transportation drew people and commerce to coasts and rivers. Thousands of years later, density of urban areas draws workers and industry seeking other benefits of agglomeration. Workers in cities have access to a greater quantity of jobs, jobs better suited for their skills, and access to productivity increasing human capital that creates wage premiums even after controlling for differences in variables like education and cost of living [85, 88, 243]. After hours, the density of cities supports cultural amenities like museums, unique restaurants, and public transportation that fuels urban migration [53, 93]. These benefits and many others scale in universal, predictable ways as cities grow in size [28, 29, 27].

This is not to say city life is problem free. Prior to proper sanitation systems, disease ran rampant and density is kindling for epidemics even today. The prosperity of cities is not distributed equally. Millions live in dangerous slums and informal settlements in the developing world while segregation based on race and income is still prevalent in many developed areas. The challenge for policy-makers, planners, and residents is to manage the growth of cities. To do that, though, they must be armed with information.

[1] http://www.mckinsey.com/insights/urbanization/us_cities_in_the_global_economy

Economic statistics are critical for decision-making by both government and private institutions. Despite their great importance, current measurements draw on limited sources of information, losing accuracy with potentially dire consequences. The beginning of the Great Recession offers a powerful case study: initial estimates by Bureau of Economic Affairs suggested that the contraction of GDP in the fourth quarter of 2008 was an annual rate 3.8%. The American Recovery and Reinvestment Act (stimulus) was passed based on this understanding in February 2009. Less than two weeks after the plan was passed, that 3.8% figure was revised to 6.2%, and subsequent revisions peg the number at a jaw dropping 8.9% – more severe than the worst quarter during the Great Depression. The government statistics were wrong and may have hampered an effective intervention. As participation rates in unemployment surveys drop, serious questions have been raised as to the declining accuracy and increased bias in unemployment numbers [128].

In this chapter we offer a methodology to infer changes in the macro economy in near real time, at arbitrarily fine spatial granularity, using data already passively collected from mobile phones. We demonstrate the reliability of these techniques by studying data from two European countries. In the first, we show it is possible to observe mass layoffs and identify the users affected by them in mobile phone records. We then track the mobility and social interactions of these affected workers and observe that job loss has a systematic dampening effect on their social and mobility behavior. Having observed an effect in the micro data, we apply our findings to the macro scale by creating corresponding features to predict unemployment rates at the province scale. In the second country, where the macro-level data is available, we show that changes in mobility and social behavior predict unemployment rates ahead of official reports and more accurately than traditional forecasts. These results demonstrate the promise of using new data to bridge the gap between micro and macro economic behaviors and track important economic indicators. Figure 4-1 shows a schematic of our methodology.

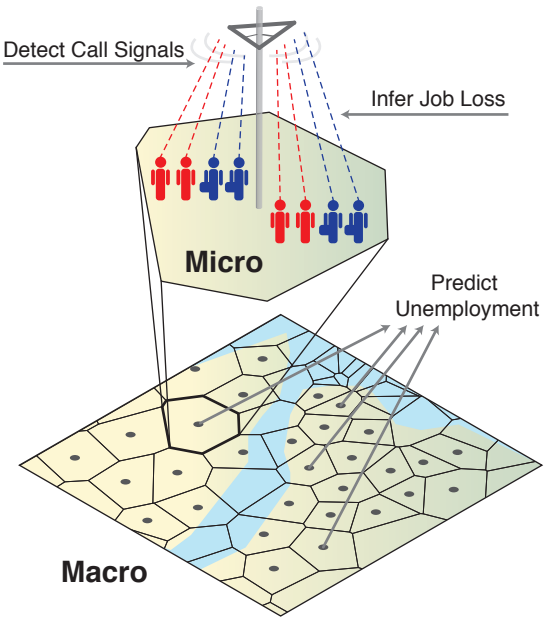

Figure 4-1: A schematic view of the relationship between job loss and call dynamics investigated at both micro and macro level.

4.1 Measuring the Economy

Contemporary macroeconomic statistics are based on a paradigm of data collection and analysis begun in the 1930s [146, 45]. Most economic statistics are constructed from either survey data or administrative records. For example, the US unemployment rate is calculated based on the monthly Current Population Survey of roughly 60,000 households, and the Bureau of Labor Statistics manually collects 80,000 prices a month to calculate inflation. Both administrative databases and surveys can be slow to collect, costly to administer, and fail to capture significant segments of the economy. These surveys can quickly face sample size limitations at fine geographies and require strong assumptions about the consistency of responses over time. Statistics inferred from survey methods have considerable uncertainty and are routinely revised in months following their release as other data is slowly collected [107, 123, 38, 128]. Moreover, changes in survey methodology can result in adjustments of reported rates of up to 1-2 percentage points [219].

The current survey-based paradigm also makes it challenging to study the effect of economic shocks on networks or behavior without reliable self-reports. This has

hampered scientific research. For example, many studies have documented the severe negative consequences of job loss in the form of difficulties in retirement [51], persistently lower wages following re-employment including even negative effects on children's outcomes [182, 168], increased risk of death and illness [216, 62], higher likelihood of divorce [52], and, unsurprisingly, negative impacts on happiness and emotional well-being [129]. Due to the cost of obtaining the necessary data, however, social scientists have been unable to directly observe the large negative impact of a layoff on the frequency and stability of an individual's social interactions or mobility.

4.2 Predicting the Present

These shortcomings raise the question as to whether existing methods could be supplemented by large-scale behavioral trace data. There have been substantial efforts to discern important population events from such data, captured by the pithy phrase "predicting the present" [59, 224, 134, 113]. Prior work has linked news stories with stock prices [108, 132, 193] and used web search or social media data to forecast labor markets [86, 11, 80, 215, 8], consumer behavior [94, 89], automobile demand, vacation destinations [59, 60]. Research on social media, search, and surfing behavior have been shown to signal emerging public health problems [91, 9, 56, 68, 95, 75, 245, 199]; although for cautionary tale see [133]. And recent efforts have even been made towards leveraging Twitter to detect and track earthquakes in real-time detection faster than seismographic sensors [185, 102, 83]. While there are nuances to the analytic approaches taken, the dominant approach has been to extract features from some large scale observational data and to evaluate the predictive (correlation) value of those features with some set of measured aggregate outcomes (such as disease prevalence). Here we offer a twist on this methodology through identification of features from observational data and to cross validate across individual and aggregate levels.

All of the applications of predicting the future are predicated in part on the presence of distinct signatures associated with the systemic event under examination. The key analytic challenge is to identify signals that (1) are observable or distinctive

enough to rise above the background din, (2) are unique or generate few false positives, (3) contain information beyond well-understood patterns such as calendar-based fluctuations, and (4) are robust to manipulation. Mobile phone data, our focus here, are particularly promising for early detection of systemic events as they combine spatial and temporal comprehensiveness, naturally incorporate mobility and social network information, and are too costly to intentionally manipulate.

Data from mobile phones has already proven extremely beneficial to understanding the everyday dynamics of social networks [19, 183, 167, 166, 58, 96] and mobility patterns of millions [15, 105, 36, 97, 206, 203, 76]. With a fundamental understanding of regular behavior, it becomes possible to explore deviations caused by collective events such as emergencies [14], natural disasters [31, 140], and cultural occasions [40, 42]. Less has been done to link these data to economic behavior. In this chapter we offer a methodology to robustly infer changes to measure employment shocks at extremely high spatial and temporal resolutions and improve critical economic indicators.

4.3 Data

We focus our analysis at three levels: the individual, the community, and the provincial levels. We begin with unemployment at the community (town) level, where we examine the behavioral traces of a large-scale layoff event. At the community and individual levels, we analyze call record data from a service provider with an approximately 15% market share in an undisclosed European country. The community-level data set spans a 15 month period between 2006 and 2007, with the exception of a gap between 6 week gap due to data extraction failures. At the province level, we examine call detail records from a service provider from another European country, with an approximately 20% market share and data running for 36 months from 2006 to 2009. Records in each data set include an anonymous id for caller and callee, the location of the tower through which the call was made, and the time the call occurred. In both cases we examine the universe of call records made over the provider's network

(see SI for more details).

4.4 Observing Unemployment at the Community Level

We study the closure of an auto-parts manufacturing plant (the plant) that occurred in December, 2006. As a result of the plant closure, roughly 1,100 workers lost their jobs in a small community (the town) of 15,000. Our approach builds on recent papers [97, 206, 203, 14] that use call record data to measure social and mobility patterns.

We model the pre-closure daily population of the town as made up of a fraction of individuals γ who will no longer make calls near the plant following its closure and the complimentary set of individual who will remain $(1-\gamma)$. As a result of the layoff, the total number of calls made near the plant will drop by an amount corresponding to the daily calls of workers who are now absent. This amounts to a structural break model that we can use to estimate the prior probability that that a user observed near the plant was laid off, the expected drop in calls that would identify them as an affected worker, and the time of the closure (see SI for full description of this model).

To verify the date of the plant closing, we sum the number of daily calls from 1955 regular users (i.e. those who make at least one call from the town each month prior to the layoff) connecting through towers geographically proximate to the affected plant. The estimator selects a break date, t_{layoff} , and pre- and post- break daily volume predictions to minimize the squared deviation of the model from the data. The estimated values are overlaid on daily call volume and the actual closure date in the Figure 4-2A. As is evident in the figure, the timing of the plant closure (as reported in newspapers and court filings) can be recovered statistically using this procedure - the optimized predictions display a sharp and significant reduction at this date. As a separate check to ensure this method is correctly identifying the break date, we estimate same model for calls from each individual user i and find distribution of these dates t^i_{layoff} is peaked around the actual layoff date (see Figure 1 in SI).

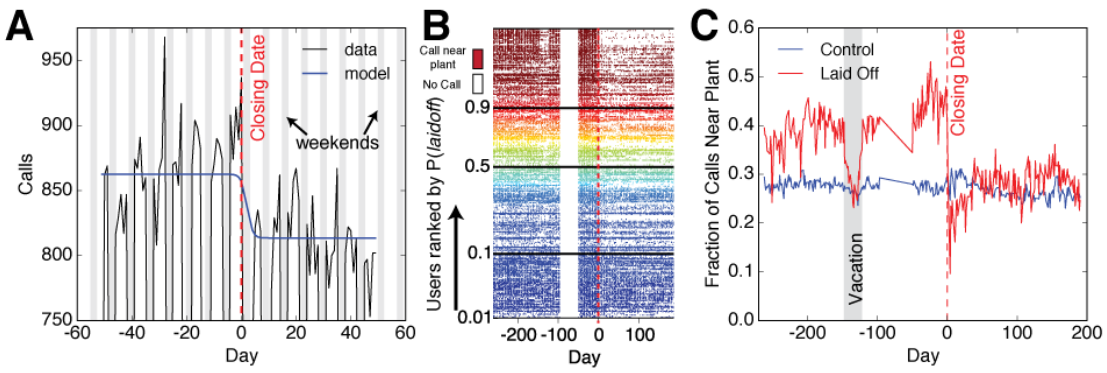

Figure 4-2: Identifying the layoff date. A) Total aggregate call volume (black line) from users who make regular calls from towers near the plant is plotted against our model (blue). The model predicts a sudden drop in aggregate call volume and correctly identifies the date of the plant closure as the one reported in newspapers and court records. B) Each of the top 300 users likely to have been laid off is represented by a row where we fill in a day as colored if a call was made near the plant on that day. White space marks the absence of calls. Rows are sorted by the assigned probability of that user being laid off according to our Bayesian model. Users with high probabilities cease making calls near the plant directly following the layoff. C) Plotting the fraction of calls made near the plant by users assigned to the top decile in probability of being unemployed (red) we see a large against a control group of individuals believed to be unaffected (blue), we see as sharp, sustained drop following the closure date. Moreover, we see that laid off individuals have an additional drop off for a two week period roughly 125 days prior the plant closure. This time period was confirmed to be a coordinated vacation for workers providing further evidence we are correctly identifying laid off workers.

4.5 Observing Unemployment at the Individual Level

To identify users directly affected by the layoff, we calculate Bayesian probability weights based on changes in mobile phone activity. For each user, we calculate the conditional probability that a user is a non-resident worker laid off as part of the plant closure based on their observed pattern of calls. To do this, we compute the difference in the fraction of days on which a user made a call near the plant in 50 days prior to the week of the layoff. We denote this difference as $\Delta q = q_{pre} - q_{post}$. We consider each user's observed difference a single realization of a random variable, Δq. Under the hypothesis that there is no change in behavior, the random variable Δq is distributed $N(0, \sqrt{\frac{q_{pre}(1-q_{pre})}{50} + \frac{q_{post}(1-q_{post})}{50}}$. Under the alternative hypothesis the individual's behavior changes pre- and post-layoff, the random variable Δq is distributed $N(d, \sqrt{\frac{q_{pre}(1-q_{pre})}{50} + \frac{q_{post}(1-q_{post})}{50}}$, where d is the mean reduction in calls

from the plant for non-resident plant workers laid off when the plant was closed. We assign user i the following probability of having been laid off given his or her calling pattern:

$$P(laidoff)_i = \frac{\gamma P(\Delta\hat{q}|\Delta q = d)}{\gamma P(\Delta\hat{q}|\Delta q = d) + (1-\gamma)P(\Delta\hat{q}|\Delta q = 0)} \tag{4.1}$$

Calculating the probabilities requires two parameters, γ, our prior that an individual is a non-resident worker at the affected plant and d, the threshold we use for the alternative hypothesis. The values of $\gamma = 5.8\%$ and $d = 0.29$ are determined based on values fit from the model in the previous section.

4.5.1 Validating the Layoff

On an individual level, Figure 4-2B shows days on which each user makes a call near the plant ranked from highest to lowest probability weight (only the top 300 users are shown, see Figure 2 in SI for more users). Users highly suspected of being laid off demonstrate a sharp decline in the number of days they make calls near the plant following the reported closure date. While we do not have ground-truth evidence that any of these mobile phone users was laid off, we find more support for our hypothesis by examining a two week period roughly 125 days prior to the plant closure. Figure 4-2C shows a sharp drop in the fraction of calls coming from this plant for users identified as laid off post closure. This period corresponds to a confirmed coordinated holiday for plant workers and statistical analysis confirms a highly significant break for individuals classified as plant workers in the layoff for this period. Given that we did not use call data from this period in our estimation of the Bayesian model, this provides strong evidence that we are correctly identifying the portion of users who were laid off by this closure. In aggregate, we assign 143 users probability weights between 50% and 100%. This represents 13% of the pre-closure plant workforce and compares closely with the roughly 15% national market share of the service provider.

4.6 Assessing the Effect of Unemployment at the Individual Level

We now turn to analyzing behavioral changes associated with job loss at the individual level. We first consider six quantities related to the monthly social behavior: A) total calls, B) number of incoming calls, C) number of outgoing calls, D) calls made to individuals physically located in the town of the plant (as a proxy for contacts made at work), E) number of unique contacts, and F) the fraction of contacts called in the previous month that were not called in the current month, referred to as churn. In addition to measuring social behavior, we also quantify changes in three metrics related to mobility: G) number of unique locations visited, H) radius of gyration, and I) average distance from most visited tower (see SI for detailed definitions of these variables). To guard against outliers such as long trips for vacation or difficulty identifying important locations due to noise, we only consider months for users where more than 5 calls were made and locations where a user recorded more than three calls.

We measure changes in these quantities using all calls made by each user (not just those near the plant) relative to months prior to the plant closure, weighting measurements by the probability an individual is laid off and relative to two reference groups: individuals who make regular calls from the town but were not believed to be laid off (mathematically we weight this group using the inverse weights from our bayesian classifier) and a random sample of 10,000 mobile phone users throughout the country (all users in this sample are weighted equally).

Figure 4-3A-I shows monthly point estimates of the average difference between relevant characteristics of users believed to be laid off compared to control groups. This figure shows an abrupt change in variables in the month directly following the plant closure. Despite this abrupt change, data at the individual level are sufficiently noisy that the monthly point estimates are not significantly different from 0 in every month. However, when data from months pre- and post-layoff are pooled, these differences are robust and statistically significant. The right panel of Figure 4-3 and

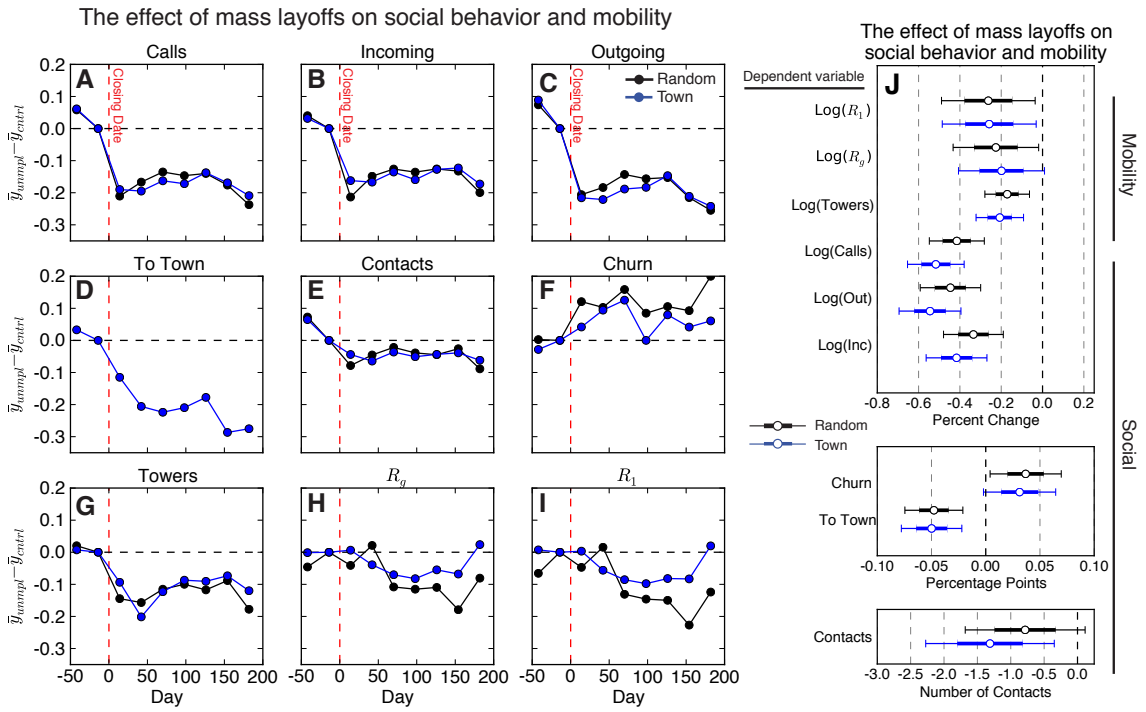

Figure 4-3: Changes in social networks and mobility following layoffs. We quantify the effect of mass layoffs relative to two control groups: users making regular calls from the town who were not identified as laid off and a random sample of users from the rest of the country. We report monthly point estimates for six social and three mobility behaviors: A) Total calls, B) number of incoming calls, C) number of outgoing calls, D) Fraction of calls to individuals in the town at the time of the call, E) number of unique contacts, and the fraction of individuals called in the previous month who were not called in the current month (churn), G) Number of unique towers visited, H) radius of gyration, I) average distance from most visited tower. Pooling months pre- and post-layoff yield statistically significant changes in monthly social network and mobility metrics following a mass layoff. J) Reports regression coefficient for each of our 9 dependent variables along with the 66% and 95% confidence intervals.

Table I in the SI show the results of OLS regressions comparing the pre-closure and post-closure periods for laid-off users relative to the two reference groups (see SI for detailed model specification as well as confidence intervals for percent changes pre- and post-layoff for each variable). The abrupt and sustained change in monthly behavior of individuals with a high probability of being laid off is compelling evidence in support of using mobile phones to detect mass layoffs with mobile phones.

We find that the total number of calls made by laid off individuals drops 51% and 41% following the layoff when compared to non-laid off residents and random users, respectively. Moreover, this drop is asymmetric. The number of outgoing calls

decreases by 54% percent compared to a 41% drop in incoming calls (using non-laid off residents as a baseline). Similarly, the number of unique contacts called in months following the closure is significantly lower for users likely to have been laid off. The fraction of calls made by a user to someone physically located in the town drops 4.7 percentage points for laid off users compared to residents of the town who were not laid off. Finally, we find that the month-to-month churn of a laid off person's social network increases roughly 3.6 percentage points relative to control groups. These results suggest that a user's social interactions see significant decline and that their networks become less stable following job loss. This loss of social connections may amplify the negative consequences associated with job loss observed in other studies.

For our mobility metrics, find that the number of unique towers visited by laid-off individuals decreases 17% and 20% relative to the random sample and town sample, respectively. Radius of gyration falls by 20% and 22% while the average distance a user is found from the most visited tower also decrease decreases by 26% relative to reference groups. These changes reflect a general decline in the mobility of individuals following job loss, another potential factor in long term consequences.

4.7 Observing Unemployment at the Province Level

The relationship between unemployment and these features of CDRs suggests a potential for predicting important, large-scale unemployment trends based on the population's call information. To perform this analysis, we use another CDR data set covering approximately 10 million subscribers in a different European country, which has been studied in prior work [167, 166, 97, 206, 203, 14]. This country experienced enormous macroeconomic disruptions, the magnitude of which varied regionally during the period in which the data are available. We supplement the CDR data set with quarterly, province-level unemployment rates from the EU Labor Force Survey, a large sample survey providing data on regional economic conditions within the EU (see SI for additional details).

We compute seven aggregated measures identified in the previous section: call

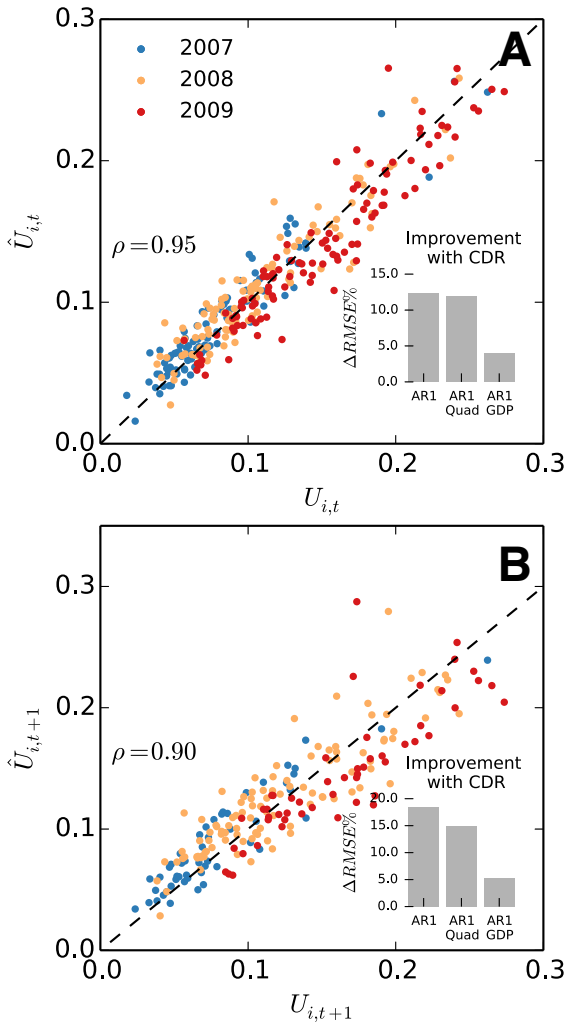

Figure 4-4: Predicting unemployment rates using mobile phone data. We demonstrate that aggregating measurements of mobile phone behaviors associated with unemployment at the individual level also predicts unemployment rates at the province level. To make our forecasts, we train various models on data from half of the provinces and use these coefficients to predict the other half. Panel A compares predictions of present unemployment rates to observed rates and Panel B shows predictions of unemployment one quarter ahead using an AR1 model that includes co-variates of behaviors measured using mobile phones. Both predictions correlate strongly with actual values while changes in rates are more difficult to predict. The insets show the percent improvement to the RMSE of predictions when mobile phone co-variates are added to various baseline model specifications. In general, the inclusion of mobile phone data reduces forecast errors by 5% to 20%.

volume, incoming calls, outgoing calls, number of contacts, churn, number of towers, and radius of gyration. Distance from home was omitted due to strong correlation with radius of gyration while calls to the town was omitted because it is not applicable

in a different country. For reasons of computational cost, we first take a random sample of 3000 mobile phone users for each province. The sample size was determined to ensure the estimation feature values are stable (see SI Figure 5 for details). We then compute the seven features aggregated per month for each individual user. The k-th feature value of user i at month t is denoted as $y_{i,t,k}$ and we compute month over month changes in this quantity as $y'_{i,t,k} = \frac{y_{i,t,k}}{y_{i,t-1,k}}$. A normalized feature value for a province s, is computed by averaging all users in selected province $\bar{y}_{s,t,k} = \sum_{i \in s} y'_{i,t,k}$. In addition, we use percentiles of the bootstrap distribution of to compute the 95% confidence interval for the estimated feature value.

After aggregating these metrics to the province level, we assess their power to improve predictions of unemployment rates. First, we correlate each aggregate measure with regional unemployment separately, finding significant correlations in the same direction as was found for individuals (see Table II in the SI). We also find the strong correlations between calling behavior variables, suggesting that principal component analysis (PCA) can reasonably be used to construct independent variables that capture changes in calling behavior while guarding against co-linearity. The first principal component, with an eigenvalue of 4.10, captures 59% of the variance in our data and is the only eigenvalue that satisfies the Kaiser criterion. The loadings in this component are strongest for social variables. Additional details on the results of PCA can be found in the SI Tables III and IV. Finally, we compute the scores for the first component for each observation and build a series of models that predict quarterly unemployment rates in provinces with and without the inclusion of this representative mobile phone variable.

First, we *predict the present* by estimating a regression of a given quarter's unemployment on calling behavior in that quarter (e.g. using phone data from Q1 to predict unemployment in Q1). As phone data is available the day a quarter ends, this method can produce predictions weeks before survey results are tabulate and released. Next, we *predict the future* in a more traditional sense by estimating a regression on a quarter's surveyed unemployment rate using mobile phone data from last quarter as a leading indicator (e.g. phone metrics from Q1 to predict unemployment rates in

Q2). This method can produce more predictions months before surveys are even conducted. See Figure 3 in the SI for a detailed timeline of data collection, release, and prediction periods. We have eight quarters of unemployment data for 52 provinces. We make and test our predictions by training our models on half of the provinces and cross-validate by testing on the other half. The groups are then switched to generate out of sample predictions for all provinces. Prediction results for an AR1 model that includes a CDR variable are plotted against actual unemployment rates in Figure 4-8. We find strong correlation coefficients between predictions of predictions of present unemployment rates ($\rho = 0.95$) as well as unemployment rates one quarter in the future ($\rho = 0.85$).

As advocated in [133] it is important to benchmark these type of prediction algorithms against standard forecasts that use existing data. Previous work has shown that the performance of most unemployment forecasts is poor and that simple linear models routinely outperform complicated non-linear approaches [160, 154, 192, 213] and the dynamic stochastic general equilibrium (DSGE) models aimed at simulating complex macro economic interactions [180, 84]. With this in mind, we compare predictions made with and without mobile phone covariates using three different model specifications: AR1, AR1 with a quadratic term (AR1 Quad), AR1 with a lagged national GDP covariate (AR1 GDP). In each of these model specifications, the coefficient related to the principal component CDR score is highly significant and negative as expected given that the loadings weigh heavily on social variables that declined after a mass layoff (see tables V and VI in the SI regression results). Moreover, adding the phone data to a to each model significantly improves forecast accuracy and reduces the root mean squared error predicting the present unemployment rate by between 5% and 20% (see inserts in Figure 4-8). As additional checks that we are capturing true improvements, we use mobile phone data from only the first half of each quarter (before surveys are even conducted) and still achieve a 3%-10% improvement in forecasts. These results hold even when variants are run to include quarterly and province level fixed effects (see tables VII and VIII in the SI).

In summary, we have shown that features associated with job loss at the individual

level are similarly correlated with province level changes in unemployment rates in a separate country. Moreover, we have demonstrated the ability of massive, passively collected data to identify salient features of economic shocks that can be scaled up to measure macro economic changes. These methods allow us to predict “present" unemployment rates two to eight weeks prior to the release of traditional estimates and predict “future" rates up to four months ahead of official reports more accurately than using historical data alone.

4.8 Discussion

We have presented algorithms capable of identifying employment shocks at the individual, community, and societal scales from mobile phone data. These findings have great practical importance, potentially facilitating the identification of macroeconomic statistics with much finer spatial granularity and faster than traditional methods of tracking the economy. We can not only improve estimates of the current state of the economy and provide predictions faster than traditional methods, but also predict future states and correct for current uncertainties. Moreover, with the quantity and richness of these data increasing daily, these results represent conservative estimates of its potential for predicting economic indicators. The ability to get this information weeks to months faster than traditional methods is extremely valuable to policy and decision makers in public and private institutions. Further, it is likely that CDR data are more robust to external manipulation and less subject to service provider algorithmic changes than most social media [133]. But, just as important, the micro nature of these data allow for the development of new empirical approaches to study the effect of economic shocks on interrelated individuals.

While this study highlights the potential of new data sources to improve forecasts of critical economic indicators, we do not view these methods as a substitute for survey based approaches. Though data quantity is increased by orders of magnitude with the collection of passively generated data from digital devices, the price of this scale is control. The researcher no longer has the ability to precisely define which

variables are collected, how they are defined, when data collection occurs making it much harder to insure data quality and integrity. In many cases, data is not collected by the researcher at all and is instead first pre-processed by the collector, introducing additional uncertainties and opportunities for contamination. Moreover, data collection itself is now conditioned on who can has specific devices and services, introducing potential biases due to economic access or sorting. If policy decisions are based solely on data derived from smartphones, the segment of the population that cannot afford these devices may be underserved.

Surveys, on the other hand, provide the researcher far more control to target specific groups, ask precise questions, and collect rich covariates. Though the expensive of creating, administering, and participating in surveys makes it difficult to collect data of the size and frequency of newer data sources, they can provide far more context about participants. This work demonstrates the benefits of both data gathering methods and shows that hybrid models offer a way to leverage the advantages of each. Traditional survey based forecasts are improved here, not replaced, by mobile phone data. Moving forward we hope to see more such hybrid approaches. Projects such as the Future Mobility Survey and the MIT Reality Mining project bridge this gap by administering surveys via mobile devices, allowing for the collection of process generated data as well as survey based data. These projects open the possibility to directly measure the correlation between data gathered by each approach.

The macro-economy is the complex concatenation of interdependent decisions of millions of individuals [130]. To have a measure of the activity of almost every individual in the economy, of their movements and their connections should transform our understanding of the modern economy. Moreover, the ubiquity of such data allows us to test our theories at scales large and small and all over the world with little added cost. We also note potential privacy and ethical issues regarding the inference of employment/unemployment at the individual level, with potentially dire consequences for individuals' access, for example, to financial markets. With the behavior of billions is being volunteered, captured, and stored at increasingly high resolutions, these data present an opportunity to shed light on some of the biggest

problems facing researchers and policy makers alike, but also represent an ethical conundrum typical of the “big data" age.

4.9 Acknowledgments

The work in this chapter was the result of a collaboration with Yu-Ru Lin, Erich Muehlegger, Daniel Shoag, Marta C. González, and David Lazer.

4.10 Chapter4 - Appendix

4.10.1 Materials and Methods

CDR Data set 1 (D1) We analyze call detail records (CDRs) from two industrialized European countries. In the first country, we obtain data on 1.95 million users from a service provider with roughly 15% market share. The data run for 15 months across the years 2006 and 2007, with the exception of a gap between August and September 2006. Each call record includes a de-identified caller and recipient IDs, the locations of the caller's and recipient's cell towers and the length of the call. Caller or recipients on other network carriers are assigned random IDs. There are approximately 1.95 million individuals identified in the data, 453 million calls, and 16 million hours of call time. The median user makes or receives 90 calls per months.

CDR Data set 2 (D2) The second data set contains 10 million users (roughly 20% market share) within a single country over three years of activity. Like D1, each billing record for voice and text services, contains the unique identifiers of the caller placing the call and the callee receiving the call, an identifier for the cellular antenna (tower) that handled the call, and the date and time when the call was placed. Coupled with a data set describing the locations (latitude and longitude) of cellular towers, we have the approximate location of the caller when placing the call. For this work we do not distinguish between voice calls

and text messages, and refer to either communication type as a "call." However, we also possess identification numbers for phones that are outside the service provider but that make or receive calls to users within the company. While we do not possess any other information about these lines, nor anything about their users or calls that are made to other numbers outside the service provider, we do have records pertaining to all calls placed to or from those ID numbers involving subscribers covered by our data set. Thus egocentric networks between users within the company and their immediate neighbors only are complete. This information was used to generate egocentric communication networks and to compute the features described in the main text. From this data set, we generate a random sample population of k users for each of the provinces, and track each user's call history during our 27-month tracking period (from December 2006 to March 2009). We discuss how the sample size is chosen in a following subsection. Finally, we note that due to an error in data extraction from the provider, we are missing data for Q4 in 2007.

4.10.2 Filtering CDR Data

We limit our sample to mobile phone users who either make or receive at least ten calls connecting through one of the five cell towers closest to the manufacturing plant of interest. In addition, we require that users make at least one call in each month spanned by a given data set to ensure users are still active.

4.10.3 Manufacturing plant closure

We gather information on a large manufacturing plant closing that affected a small community within the service provider's territory from news articles and administrative sources collected by the country's labor statistics bureau. The plant closure occurred in December 2006 and involved 1,100 employees at an auto-parts manufacturing plant in a town of roughly 15,000 people.

4.10.4 Town Level Structural Break Model

We model the pre-closure daily population of the town as consisting of three segments: a fraction of non-resident plant workers γ, a fraction of resident workers μ, and a fraction of non-workers normalized to $(1-\gamma-\mu)$. We postulate that each individual i has a flow probability of making or receiving a call at every moment p_i. Workers spend a fraction ψ of their day at their jobs and thus make, in expectation, $p_i\psi$ call on a given day during work hours. When losing their job in the town, both resident and non-resident workers are re-matched in national, not local, labor market.

Given this model, the expected daily number of cell phone subscribers making or receiving calls near the plant is:

$$vol = \begin{cases} \gamma\psi\bar{p} + (1-\gamma)\bar{p} & \text{for } t < t_{layoff} \\ \mu(1-\psi)\bar{p} + (1-\gamma-\mu)\bar{p} & \text{for } t \geq t_{layoff} \end{cases}$$

This model predicts a discrete break in daily volume from the towers proximate to the plant of $(\gamma+\mu)\psi\bar{p}$ at the date t_{layoff}. For workers, the predicted percentage change in call volume from these towers is $\frac{(\gamma+\mu)\psi\bar{p}}{(\mu\bar{p}+\gamma\psi\bar{p}}$. Non-workers experience no change.

4.10.5 Individual Structural Break Model

We fit a model similar to the community structural break model to data for each individual user, i, based on the probability they made a call from the town on each day. For each individual, we use the non-linear estimator to select a break date t^i_{layoff} , and constant pre- and post- break daily probabilities $p_{i,t<t^i_{layoff}}$ and $p_{i,t>t^i_{layoff}}$ to minimize the squared deviation from each individuals' data. Figure 4-5 plots the distribution of break-dates for individuals. As expected, there is a statistically significant spike in the number of individuals experiencing a break in the probability of making a call from the town at the time of the closure and significantly fewer breaks on other, placebo dates. These two methods provide independent, yet complementary ways of detecting mass layoffs in mobile phone data.

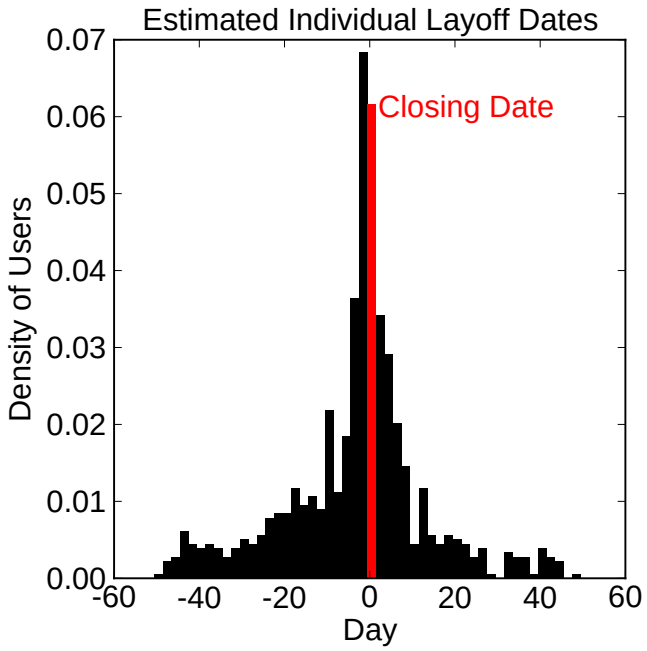

Figure 4-5: We plot the distribution of break dates for the structural break model estimated for individuals. We find a strong, statistically significant peak centered on the reported closure date (red) with far fewer breaks on other, placebo dates. This is consistent with both our community wide model as well as the Bayesian model presented above.

4.10.6 Bayesian Estimation

On an individual level, Figure 4-6A shows days on which each user makes a call near the plant ranked from highest to lowest probability weight. Figure 4-6B provides greater detail for users probability weights between 50% and 100%. Users highly suspected of being laid off demonstrate a sharp decline in the number of days they make calls near the plant following the reported closure date. Figure 4-6C graphs the inverse cumulative distribution of probability weights. While we do not have ground-truth evidence that any of these mobile phone users was laid off, we find more support for our hypothesis by examining a two week period roughly 125 days prior to the plant closure.

Figures 4-6A and 4-6B illustrate that the call patterns of users assigned the highest probabilities change significantly after the plant closure. These users make calls from the town on a consistent basis before the layoff, but make significantly fewer calls from the town afterwards. In contrast, the call patterns of users assigned the lowest weights do not change following the plant closure. In aggregate, we assign 143 users probability weights between 50% and 100%. This represents 13% of the pre-closure plant workforce ĂŞ this fraction compares closely with the roughly 15% national market share of the service provider.

4.10.7 The European Labor Force Survey

Each quarter, many European countries are required to conduct labor force surveys to measure important economic indicators like the unemployment rates studied here. In person or telephone interviews concerning employment status are conducted on a sample size of less than 0.5% of the population. Moreover, participants are only asked to provide responses about their employment status during a 1 week period in the quarter.

These “microdata" surveys are then aggregated at the province and national levels. Confirmed labor force reports and statistics for a particular quarter are released roughly 14 weeks after the quarter has ended. For example, Q1 of 2012 begins January

1st, 2012 and ends March 31st, 2012. The survey data is analyzed and unemployment numbers are released between two and three weeks following the end of the quarter. These numbers, however, are unconfirmed and subject to revisions which can occur at any time in the following quarters.

4.10.8 The Effect of Job Loss on Call Volumes

We measure the effect of job loss on six properties of an individual's social behavior and three mobility metrics.

CDR Metrics

Calls: The total number of calls made and received by a user in a given month.

Incoming: The number of calls received by a user in a given month.

Outgoing: The number of calls made by a user in a given month

Contacts: The number of unique individuals contacted by a user each month. Includes calls made and received.

To Town: The fraction of a user's calls made each month to another user who is physically located in the town of the plant closure at the time the call was made.

Churn: The fraction of a user's contacts called in the previous month that was not called in the current month. Let C_t be the set of users called in month t. Churn is then calculated as: $churn_t = 1 - \frac{|C_{t-1} - C_t|}{|C_t - 1|}$.

Towers: The number of unique towers visited by a user each month.

Radius of Gyration, R_g: The average displacement of a user from his or her center of mass: $R_g = \sqrt{\frac{1}{n}\sum_{j=1}^{n} |\vec{r}_j - \vec{r}_{cm}|^2}$, where n is the number of calls made by a user in the month and r_{cm} is the center of mass calculated by averaging the positions of all a users calls that month.

Average Distance from Top Tower, R_1: The average displacement of a user from their most called location: $R_1 = \sqrt{\frac{1}{n}\sum_{j=1}^{n} |\vec{r}_j - \vec{r}_1|^2}$, where n is the number of calls made by a user in the month and r_1 is the coordinates of the location most visited by the user.

4.10.9 Measuring Changes

For each user i, we compute these metrics monthly. Because individuals may have different baseline behaviors, we normalize a user's time series to the month immediately before the layoff denoted t^*. To assess differences in behavior as a result of the mass layoff, we construct three groups: (1) A group of laid off users from the town where the probability of being laid off is that calculated in the previous section, (2) a town control group consisting of the same users as group 1, but with inverse weights, and (3) a group of users selected at random from the country population. Each user in the final group is weighted equally.

For each month, we compute the weighted average of all metrics then plot the difference between the laid off group and both control groups in Figure 4-3.

$$y_t = \sum_i w_i \frac{y_{i,t}}{y_{i,t^*}} \tag{4.2}$$

$$\Delta y_t = \bar{y}_t - \bar{y}_{t,control} \tag{4.3}$$

We estimate changes in monthly behavior using OLS regressions. We specify two models that provide similar results. For a metric :

$$y_i = \alpha_i + \beta_1 A_i + \beta_2 U_i + \beta_3 A_i U_i \tag{4.4}$$

where A_i is a dummy variable indicating if the observation was made in a month before or after the plant closure and U_i is a dummy variable that is 1 if the user was assigned a greater than 60% probability of having been laid off and 0 otherwise. An alternate model substitutes the probability of layoff itself, for the unemployed

dummy:

$$y_i = \alpha_i + \beta_1 A_i + \beta_2 w_i + \beta_3 A_i w_i \tag{4.5}$$

In many cases, we are more interested in relative changes in behavior rather than absolute levels. For this, we specify a log-level model of the form:

$$\log(y_i) = \alpha_i + \phi_1 A_i + \phi_2 w_i + \phi_3 A_i w_i \tag{4.6}$$

Now the coefficient ϕ_3 can be interpreted as the percentage change in feature $y_{i,n}$ experienced by a laid off individual in months following the plant closure. Changes to mobility metrics as well as changes to total, incoming, and outgoing calls were estimated using the log-level model. Churn and To Town metrics are percentages already and are thus estimated using a level-level model. The changes in the number of contacts each also estimated using a level-level model.

Models are estimated using data from users believed to be unemployed and data from the two control groups. Results are shown in Table 4.1. Comparisons to each group produce consistent results.

4.10.10 Predicting Province Level Unemployment

To evaluate the predictive power of micro-level behavioral changes, we use data from a different undisclosed industrialized European country. As discussed in the main text, we use call detail records spanning nearly 3 years and the entire user base of a major mobile phone provider in the country. For each of the roughly 50 provinces within this country, we assemble quarterly unemployment rates during the period covered by the CDR data. At the national level, we collect a time series of GDP. We select a sample of users in each province and measure the average relative value of 7 of the variables identified to change following a layoff. To-town and distance from home variables are omitted as the former is only measured when we know the location of the layoff and the latter is strongly correlated with R_g.

First, we correlate each aggregate calling variable with unemployment at the re-

gional level. To control for differences in base levels of unemployment across the country, we first de-mean unemployment and each aggregate variable. Table 4.2 shows that each calling behavior is significantly correlated with unemployment and that these correlations are consistent with the directions found in the individual section of the chapter. Moreover, we discover strong correlation between each of the calling behavior variables, suggesting that principal component analysis is appropriate.

Principal Component analysis

As shown in the individual section of the chapter, changes in these variables following a mass layoff are correlated. This correlation is seen in province level changes as well (Table 4.2). Given this correlation, we use principal component analysis (PCA) extract an independent mobile phone variable and guard against co-linearity when including all phone variables as regressors. The results from PCA and the loadings in each component can be found in Table 4.3 and Table 4.4, respectively. We find only the first principal component passes the Kaiser test with an eigenvalue significantly greater than 1, but that this component captures 59% of the variance in the calling data. The loadings in this component fall strongly on the social variables behavior. We then compute the scores for this component for each observation in the data and use these scores as regressors. The prominent elements of the first principal component are primarily related to the social behavior of callers.

Model Specification

We make predictions of present and future unemployment rates using three different models specifications of unemployment where each specification is run in two variants, one with the principal component score as an additional independent variable denoted as CDR_t and the other without. The sixteen models are described as follows:

1. **AR(1)**

$$U_t = \alpha_1 U_{t-1} \tag{4.7}$$

$$U_t = \beta_1 U_{t-1} + \gamma CDR_t \tag{4.8}$$

$$U_{t+1} = \alpha_1 U_{t-1} \tag{4.9}$$

$$U_{t+1} = \beta_1 U_{t-1} + \gamma CDR_t \tag{4.10}$$

2. **AR(1) + Quad**

$$U_t = \alpha_1 U_{t-1} + \alpha_2 U_{t-1}^2 \tag{4.11}$$

$$U_t = \beta_1 U_{t-1} + \beta_2 U_{t-1}^2 + \gamma CDR_t \tag{4.12}$$

$$U_{t+1} = \alpha_1 U_{t-1} + \alpha_2 U_{t-1}^2 \tag{4.13}$$

$$U_{t+1} = \beta_1 U_{t-1} + \beta_2 U_{t-1}^2 + \gamma CDR_t \tag{4.14}$$

3. **AR(1) + GDP**

$$U_t = \alpha_1 U_{t-1} + \alpha_2 gdp_{t-1} \tag{4.15}$$

$$U_t = \beta_1 U_{t-1} + \beta_2 gdp_{t-1} + \gamma CDR_t \tag{4.16}$$

$$U_{t+1} = \alpha_1 U_{t-1} + \alpha_2 gdp_{t-1} \tag{4.17}$$

$$U_{t+1} = \beta_1 U_{t-1} + \beta_2 gdp_{t-1} + \gamma CDR_t \tag{4.18}$$

To evaluate the ability of these models to predict unemployment, we use a cross-validation framework. Data from half of the provinces are used to train the model and these coefficients are used to predict unemployment rates given data for the other half of the provinces. We perform the same procedure switching the training and testing set and combine the out of sample predictions for each case. We evaluate the overall utility of these models by plotting predictions versus observations, finding strong correlation (see the main text). To evaluate the additional benefit gained from the inclusion of phone data, we compute the percentage difference between the

same model specification with and without the mobile phone data, $\Delta RMSE\% = 1 - RMSE_{w/CDR}/RMSE_{w/out}$ .In each case, we find that the addition of mobile phone data reduces the RMSE by 5% to 20%.

Predictions using weekly CDR Data

Until now, we have have used data from the entire quarter to predict the results from the unemployment survey conducted in the same quarter. While these predictions would be available at the very end of the quarter, weeks before the survey data is released, we also make predictions using CDR data from half of each quarter to provide an additional 1.5 months lead time that may increase the utility of these predictions. We estimate the same models as described in the previous section and find similar results. Even without full access to a quarter's CDR data, we can improve predictions of that quarter's unemployment survey before the quarter is over by 3%-6%.

The Effect of Sample Size on Feature Estimation

It is important to consider the extent to which the sampling size is sufficient and does not affect much the feature estimation. We study the reliability of sample size (k) in terms of relative standard deviation (RSD). For each given sample size k, we sample T times (without replacement) from the population. The RSD with respect to sample size k for a particular feature, is given by $RSD(k) = \frac{s_k}{f_k}$ where s_k is the standard deviation of the feature estimates from the T samples, and f_k is the mean of the feature estimates from the T samples. We use $T = 10$ to study the feature reliability. In Figure 4-9, we plot the different features' %RSD by averaging the RSD values of all provinces. The plots show that the values of %RSD over sample size $k = 100, 200, ..., 2000$ decrease rapidly. When sample size $k = 2000$, the %RSD for all features, except for radius of gyration (R_g), is lower than 1%. The estimates of R_g exhibit the highest variation; however, we can still obtain reliable estimates with thousands of sampled individuals ($RSD(k) = 0.026$ for R_g, with $k = 2000$).

4.10.11 Mass Layoffs and General Unemployment

While mass layoffs provide a convenient and interesting natural experiment to deploy our methods, they are only one of many employment shocks that economy absorbs each month. We have measured changes in call behaviors due to mass layoffs, but these changes may be unhelpful if they do not result from other forms of unemployment like isolated layoffs of individual works. Though it is beyond the scope of this work to directly determine if individuals affected by mass layoffs experience the same behavioral changes as those experiencing unemployment due to other reasons, we do find strong correlations between the number of mass layoffs observed in a given time period and general unemployment rates.

Using monthly data provided by the United States Bureau of Labor Statistics (BLS), Figure 4-10 shows time series of the number of monthly initial claimants of unemployment benefits do to any change in employment status and due to mass layoffs directly. There is similarly high correlation between the number of distinct mass layoff events (irrespective of the number of claimants in each event). While the relationship between claimants due to mass layoffs and the overall unemployment rate is not as strong, there is still significant correlation (see Table 4.9). Moreover, these positive correlations hold true at a state level as shown by Table 4.10.

Table 4.1: Regression Results - Social and Mobility Measures.

	(1) Log(calls)	(2) Log(inc)	(3) Log(out)	(4) contacts	(5) to town	(6) churn	(7) Log(towers)	(8) Log(R_g	(9) Log(to top)
Panel A: Compared to Random User									
Post-Layoff Dummy	0.0390**	0.0464***	0.0448**	0.213***	-0.00000264	0.0201***	0.0179	0.0657*	0.0588*
	(0.0155)	(0.0170)	(0.0193)	(0.0124)	(0.000135)	(0.00592)	(0.0721)	(0.0337)	(0.0352)
Laidoff Dummy * Post-Layoff Months	-0.415***	-0.335***	-0.446***	-0.785*	-0.0478***	0.0368**	-0.171***	-0.226**	-0.262**
	(0.0679)	(0.0738)	(0.0747)	(0.458)	(0.0136)	(0.0167)	(0.0550)	(0.106)	(0.116)
Observations	10011	9742	9456	10011	10011	10011	10011	6922	6908
R-Squared	0.828	0.805	0.803	0.923	0.892	0.338	0.812	0.655	0.657
Panel B: Comparison to Non-laidoff Town Users									
Post-Layoff Dummy	0.0511***	0.0497***	0.0574***	0.454***	-0.000999	0.0301***	0.00973	0.0345**	0.0371**
	(0.0106)	(0.0122)	(0.0118)	(0.110)	(0.00188)	(0.00368)	(0.00935)	(0.0166)	(0.0175)
Laidoff Dummy * Post-Layoff Months	-0.517***	-0.416***	-0.545***	-1.311***	-0.0499***	0.0312*	-0.207***	-0.199*	-0.258**
	(0.0679)	(0.0738)	(0.0747)	(0.458)	(0.0136)	(0.0167)	(0.0550)	(0.106)	(0.116)
Observations	17506	17342	17118	17506	17506	17506	17506	15474	15417
R-Squared	0.875	0.860	0.871	0.938	0.889	0.349	0.899	0.729	0.741

All specifications include user fixed effects. Robust standard errors clustered by user.
*, **, and *** denote significance at the 10%, 5% and 1% levels.

Table 4.2: Correlation coefficients between normalized, aggregated calling behaviors and unemployment rates the province level.

	(1)							
	Calls	Inc	Out	Contacts	Churn	Towers	R_g	Unemp
Calls	1							
Inc	0.746***	1						
Out	0.952***	0.724***	1					
Contacts	0.807***	0.667***	0.858***	1				
Churn	0.102*	-0.141**	0.132**	0.183***	1			
Towers	0.701***	0.522***	0.711***	0.584***	0.196***	1		
R_g	0.373***	0.193***	0.379***	0.269***	0.177***	0.696***	1	
Unemp	-0.428***	-0.356***	-0.396***	-0.169***	0.138**	-0.418***	-0.295***	1

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4.3: PCA results for call variables.

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	4.10	2.93	0.59	0.59
Comp2	1.17	0.31	0.17	0.75
Comp3	0.86	0.53	0.12	0.89
Comp4	0.34	0.04	0.05	0.93
Comp5	0.29	0.11	0.04	0.97
Comp6	0.19	0.15	0.03	0.99
Comp7	0.04	.	0.01	1.00

Table 4.4: PCA Loadings. Significant elements are bolded.

	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7
Calls	**0.47**	-0.117	0.098	0.013	-0.147	**-0.568**	**0.643**
Inc	**0.39**	**-0.376**	-0.020	0.232	**0.794**	0.131	-0.054
Out	**0.47**	-0.095	0.142	0.008	-0.268	**-0.346**	**-0.746**
Contacts	**0.431**	-0.096	0.284	0.167	**-0.406**	**0.710**	0.156
Churn	0.10	0.726	**0.605**	0.073	0.298	-0.046	-0.002
Towers	**0.388**	0.247	**-0.308**	**-0.802**	0.125	0.179	0.015
R_g	0.252	**0.487**	**-0.652**	**0.517**	-0.065	0.015	-0.006

Table 4.5: Predicting Present Unemployment Rates - Cross Validation Model Coefficients

	Model Estimates with out CDR Data					
	(1) AR1	(2) AR1	(3) AR1 Quad	(4) AR1 Quad	(5) AR1 GDP	(6) AR1 GDP
U_{t-1}	1.123***	1.061***	1.195***	1.599***	1.072***	1.032***
	(0.0325)	(0.0456)	(0.152)	(0.139)	(0.0316)	(0.0441)
U_{t-1}^2			-0.301	-2.001***		
			(0.615)	(0.562)		
gdp_{t-1}					0.00119***	0.00103***
					(0.000169)	(0.000149)
_cons	-0.00180	0.00304	-0.00541	-0.0256***	-0.0244***	-0.0183***
	(0.00343)	(0.00407)	(0.00788)	(0.00748)	(0.00423)	(0.00430)
Observations	144	168	144	168	144	168
R^2	0.871	0.891	0.872	0.903	0.898	0.910
Adjusted R^2	0.871	0.891	0.870	0.902	0.897	0.909

	Model Estimates with CDR Data					
	(1) AR1	(2) AR1	(3) AR1 Quad	(4) AR1 Quad	(5) AR1 GDP	(6) AR1 GDP
U_{t-1}	1.064***	1.076***	1.180***	1.395***	1.056***	1.062***
	(0.0330)	(0.0389)	(0.146)	(0.139)	(0.0323)	(0.0403)
CDR_t	-0.00597***	-0.00605***	-0.00602***	-0.00534***	-0.00313**	-0.00481***
	(0.000928)	(0.000842)	(0.000925)	(0.000917)	(0.00120)	(0.00120)
U_{t-1}^2			-0.486	-1.192*		
			(0.586)	(0.566)		
gdp_{t-1}					0.000855***	0.000390
					(0.000225)	(0.000211)
_cons	0.00434	0.00290	-0.00144	-0.0141	-0.0148*	-0.00516
	(0.00338)	(0.00350)	(0.00767)	(0.00750)	(0.00591)	(0.00539)
Observations	144	168	144	168	144	168
R^2	0.893	0.918	0.894	0.922	0.902	0.920
Adjusted R^2	0.892	0.917	0.891	0.920	0.900	0.918

Standard errors in parentheses
Coefficients are reported for models trained on each half of the data.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4.6: Predicting Future Unemployment Rates - Cross Validation Model Coefficients

	Model Estimates with out CDR Data					
	(1) AR1	(2) AR1	(3) AR1 Quad	(4) AR1 Quad	(5) AR1 GDP	(6) AR1 GDP
U_{t-1}	1.161***	1.040***	1.351***	2.140***	1.074***	0.991***
	(0.0669)	(0.0871)	(0.329)	(0.268)	(0.0601)	(0.0839)
U_{t-1}^2			-0.852	-4.051***		
			(1.421)	(0.983)		
gdp_{t-1}					0.00220***	0.00194***
					(0.000213)	(0.000174)
_cons	0.00906	0.0164*	-0.0000247	-0.0428**	-0.0349***	-0.0256**
	(0.00730)	(0.00776)	(0.0165)	(0.0150)	(0.00668)	(0.00801)
Observations	96	112	96	112	96	112
R^2	0.703	0.730	0.704	0.773	0.813	0.812
Adjusted R^2	0.700	0.728	0.697	0.769	0.809	0.808
	Model Estimates with CDR Data					
	(1) AR1	(2) AR1	(3) AR1 Quad	(4) AR1 Quad	(5) AR1 GDP	(6) AR1 GDP
U_{t-1}	1.074***	1.135***	1.490***	1.877***	1.058***	1.070***
	(0.0608)	(0.103)	(0.272)	(0.300)	(0.0583)	(0.114)
CDR_t	-0.0137***	-0.0117***	-0.0140***	-0.0102***	-0.00716**	-0.00697*
	(0.00159)	(0.00180)	(0.00154)	(0.00211)	(0.00235)	(0.00309)
U_{t-1}^2			-1.870	-2.775*		
			(1.229)	(1.314)		
gdp_{t-1}					0.00147***	0.00108**
					(0.000324)	(0.000380)
_cons	0.0208***	0.0122	0.00110	-0.0278	-0.0141	-0.00932
	(0.00598)	(0.00819)	(0.0130)	(0.0155)	(0.00965)	(0.00655)
Observations	96	112	96	112	96	112
R^2	0.801	0.814	0.805	0.833	0.828	0.825
Adjusted R^2	0.797	0.811	0.798	0.828	0.822	0.821

Standard errors in parentheses
Coefficients are reported for models trained on each half of the data.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4.7: RMSE with the addition of CDR data from the entire quarter with various fixed effects. The best performing model is bolded.

	No F.E.			Quarter F.E.			Province F.E.		
Model	RMSE	w/ CDR	ΔRMSE%	RMSE	w/ CDR	ΔRMSE%	RMSE	w/ CDR	ΔRMSE%
Present									
AR1	0.0192	0.0168	12.37	0.195	0.0179	8.18	0.0217	0.0171	21.34
AR1 Quad	0.0191	0.0168	11.88	0.0190	0.0178	6.23	0.0218	0.0171	21.18
AR1 GDP	0.0173	**0.0166**	3.93	0.0178	0.0176	0.89	0.0179	0.0168	6.24
Future									
AR1	0.0315	0.0257	18.45	0.0315	0.0281	10.76	0.0365	0.0318	12.92
AR1 Quad	0.0308	0.0262	14.89	0.0302	0.0285	5.55	0.0358	0.0314	12.21
AR1 GDP	0.0260	**0.0246**	5.37	0.0270	0.0269	0.61	0.0274	0.0313	-14.12

Table 4.8: RMSE with the addition of CDR data from the half of the quarter with various fixed effects. The best performing model is bolded.

	No F.E.			Quarter F.E.			Province F.E.		
Model	RMSE	w/ CDR	ΔRMSE%	RMSE	w/ CDR	ΔRMSE%	RMSE	w/ CDR	ΔRMSE%
Present									
AR1	0.0192	0.0164	14.04	0.0194	0.0172	11.12	0.0217	0.0165	23.82
AR1 Quad	0.0190	**0.0163**	14.19	0.0189	0.0170	10.01	0.0217	0.0164	24.58
AR1 GDP	0.0173	0.0165	4.49	0.0177	0.0175	1.08	0.0179	0.0165	7.51
Future									
AR1	0.0314	0.0258	17.80	0.0315	0.0273	13.07	0.0365	0.0267	26.73936
AR1 Quad	0.0307	**0.0246**	19.77	0.0301	0.0262	13.02	0.0357	0.0256	28.36
AR1 GDP	0.0259	0.0258	0.50	0.0270	0.0270	-0.046	0.0274	0.0272	0.49

Table 4.9: Correlations between general unemployment and unemployment resulting from mass layoffs.

	Correlation Coefficient
Mass Layoff Events vs. Total Initial Unemployment Claimants	0.86
Initial Claimants due to Mass Layoffs vs. Total Initial Unemployment Claimants	0.73
Initial Claimants due to Mass Layoffs vs. Unemployment Rate	0.46

Table 4.10: Correlations between mass layoff and unemployment and the state level.

State	Mass Layoff Events vs. Unemp. Rate	Mass Layoff Events vs. Unemp. Claims	Mass Layoff Claims vs. Unemp. Rate	Mass Layoff Claims vs. Unemp. Claims
Alabama	0.27	0.27	0.24	0.24
Alaska	0.13	0.23	0.04	0.17
Arizona	0.36	0.35	0.21	0.21
Arkansas	0.37	0.38	0.25	0.27
California	0.44	0.42	0.30	0.27
Colorado	0.40	0.39	0.35	0.33
Connecticut	0.27	0.28	0.15	0.16
Delaware	0.41	0.40	-0.22	-0.23
District of Columbia	-0.38	-0.36	-0.46	-0.50
Florida	0.69	0.70	0.67	0.67
Georgia	0.42	0.42	0.30	0.30
Hawaii	0.46	0.46	0.37	0.36
Idaho	0.31	0.31	0.32	0.28
Illinois	0.48	0.49	0.50	0.51
Indiana	0.33	0.34	0.19	0.20
Iowa	0.23	0.24	0.18	0.19
Kansas	0.33	0.34	0.25	0.25
Kentucky	0.48	0.48	0.31	0.32
Louisiana	0.44	0.45	0.43	0.43
Maine	0.08	0.08	0.06	0.05
Maryland	0.10	0.12	0.03	0.04
Massachusetts	0.12	0.12	0.09	0.09
Michigan	0.16	0.16	0.09	0.09
Minnesota	0.25	0.25	0.10	0.10
Mississippi	0.29	0.30	0.26	0.28
Missouri	0.31	0.31	0.12	0.11
Montana	0.34	0.38	0.26	0.27
Nebraska	0.15	0.13	0.11	0.10
Nevada	0.65	0.64	0.43	0.41
New Hampshire	0.30	0.28	0.11	0.08
New Jersey	0.28	0.28	0.17	0.17
New Mexico	0.25	0.32	0.08	0.16
New York	0.34	0.36	0.28	0.29
North Carolina	0.51	0.47	0.44	0.41
North Dakota	0.30	0.31	0.10	0.12
Ohio	0.25	0.25	0.15	0.15
Oklahoma	0.27	0.27	0.11	0.10
Oregon	0.39	0.40	0.33	0.33
Pennsylvania	0.40	0.41	0.33	0.34
Puerto Rico	0.17	0.22	0.01	0.05
Rhode Island	-0.07	-0.08	-0.19	-0.19
South Carolina	0.29	0.27	0.11	0.08
South Dakota	0.60	0.58	0.15	0.12
Tennessee	0.42	0.43	0.35	0.36
Texas	0.51	0.45	0.39	0.32
Utah	0.38	0.43	0.34	0.37
Vermont	0.31	0.32	0.22	0.23
Virginia	0.44	0.39	0.20	0.15
Washington	0.42	0.42	0.39	0.35
West Virginia	0.37	0.38	0.26	0.27
Wisconsin	0.34	0.34	0.27	0.26
Wyoming	-	-	-	-

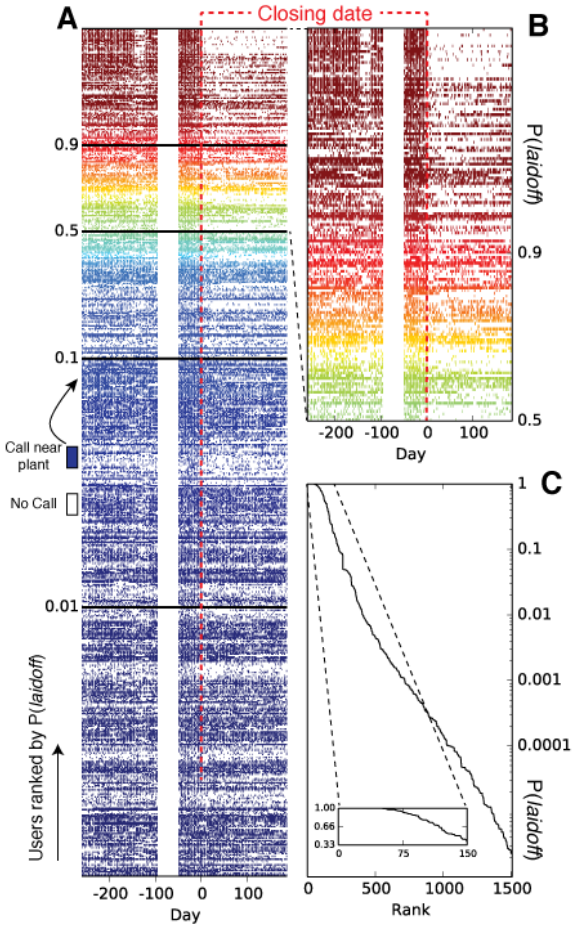

Figure 4-6: Identifying affected individuals. A) Each user is represented by a row where we fill in a day as colored if a call was made near the plant on that day. White space marks the absence of calls. Rows are sorted by the assigned probability of that user being laid off. B) A closer view of the users identified as mostly to have been laid off reveals a sharp cut off in days on which calls were made from the plant. C) An inverse cumulative distribution of assigned probability weights. The insert shows an enlarged view at the probability distribution for the 150 individuals deemed most likely to have been laid off.

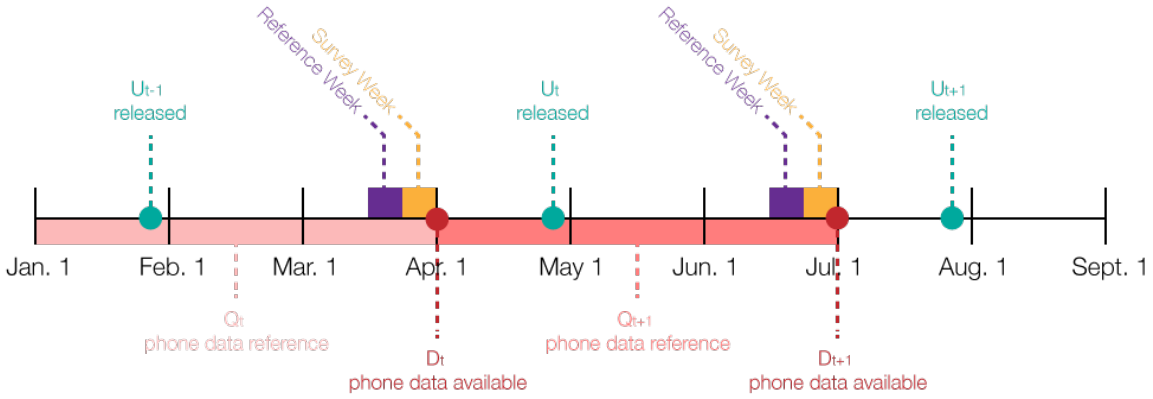

Figure 4-7: A timeline showing the various data collection and reporting periods. Traditional survey method perform surveys over the course of a single week per quarter, asking participants about their employment status during a single reference week. Unofficial survey results, subject to revision are then released a few weeks following the end of the quarter. Mobile phone data, however, is continually collected throughout the quarter and is available for analysis at any time during the period. Analysis of a given quarter can be performed and made available immediately following the end of the month.

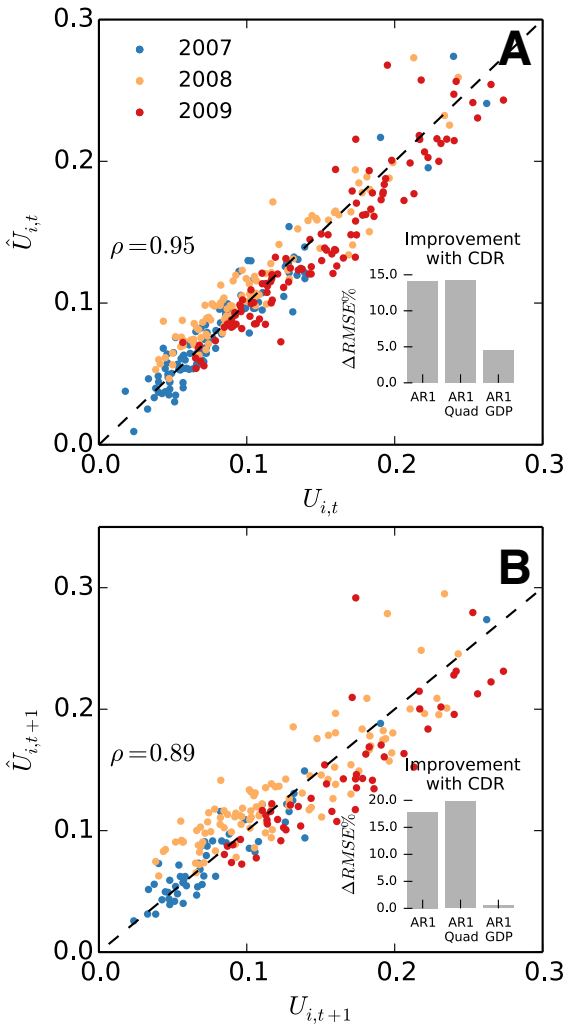

Figure 4-8: Predicting unemployment rates using mobile phone data from only the first 6 weeks of each quarter. We follows the same procedure as the main text. Panel A compares predictions of present unemployment rates to observed rates and Panel B shows predictions of unemployment one quarter ahead using a simple AR1 model that includes co-variates of behaviors measured using mobile phones. Both predictions correlate strongly with actual values while changes in rates are more difficult to predict. The insets show the percent improvement to the RMSE of predictions when mobile phone co-variates are added to each of four traditional forecasting models. In general, mobile phone data reduces forecast errors by 3% to 6%.

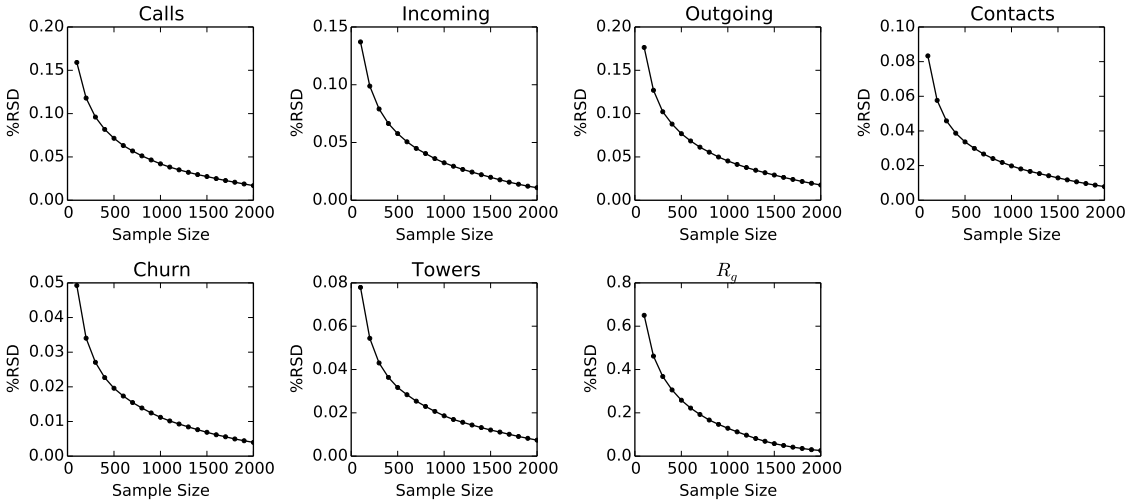

Figure 4-9: The average values of %RSD against the number of samples per province for different features. For all features, the %RSD's decrease rapidly with sample size and stabilize to relative small values before $k = 2000$.

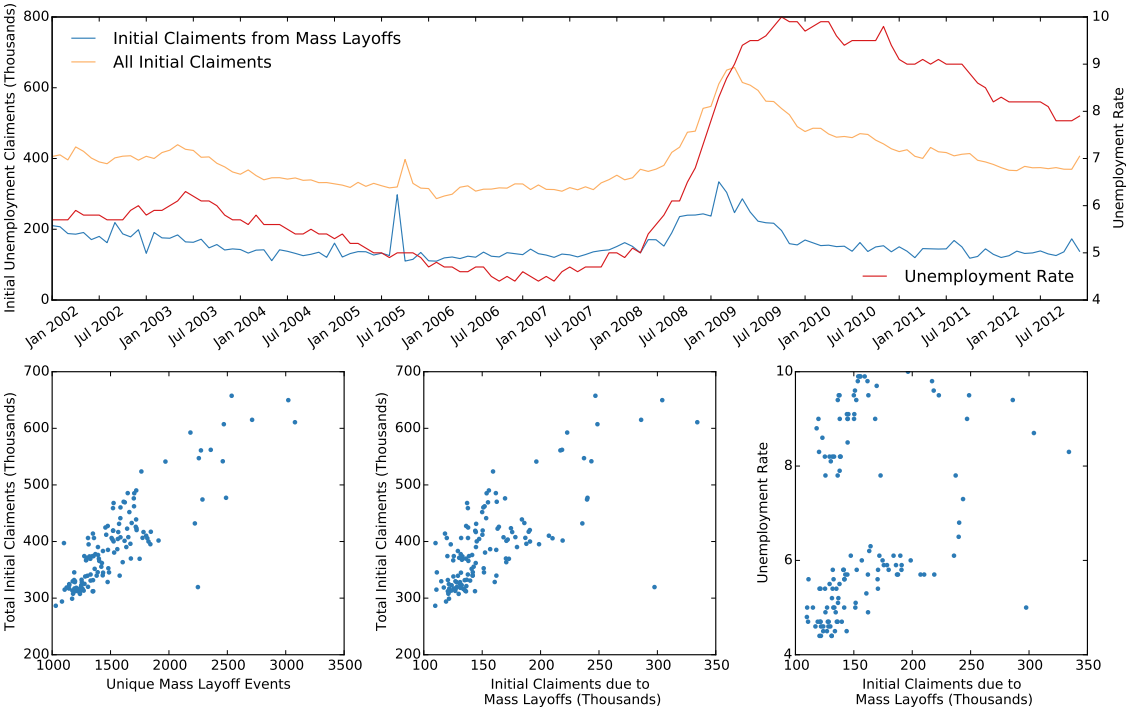

Figure 4-10: Correlations between mass layoff events and general unemployment. Using BLS data, we plot various correlations between the number of mass layoff events, the number of initial unemployment claimants due to these events, general unemployment claims, and the unemployment rate. We find strong correlation between all of these variables suggesting that mass layoffs are a good proxy for general unemployment shocks, at least at the predictive level.

Chapter 5

The Path Most Traveled : Estimating Travel Demand With Big Data Resources

5.1 Introduction

task for transportation and urban planners. To meet these challenges in the past, methods such as the widely used four-step model and more recent activity based models were developed to make use of available data computational resources. These models combine meticulous methods of statistical sampling in local [69, 201] and national household travel surveys [214, 178] to process and infer trip information between areas of a city. The estimates they produce are critically important for understanding the use of transportation infrastructure and planning for for its future[223, 210, 144, 138, 111, 110, 109, 139, 46, 24].

While the surveys that provide the empirical foundation for these models offer a combination of highly detailed travel logs for carefully selected representative population samples, they are expensive to administer and participate in. As a result, the time between surveys range from 5 to 10 years in even the most developed cities. The rise of ubiquitous mobile computing has lead to a dramatic increase in new, *big*

data resources that capture the movement of vehicles and people in near real time and promise solutions to some of these deficiencies. With these new opportunities, however, come new challenges of estimation, integration, and validation with existing models. While these data are available nearly instantaneously and provide large, long running, samples at low cost, they often lack important contextual demographic information due to privacy reasons, lack resolution to infer choices of mode, and have their own noise and biases that must be accounted for. Despite these issues, their use for urban and transportation planning has the potential to radically decrease the time in-between updated surveys, increase survey coverage, and reduce data acquisition costs. In order to realize these benefits, a number of challenges must be overcome to integrate new data sources into traditional modeling and estimation tools.

Analyzed on its own, data generated by the pervasive use of cellular phones has offered insights into abstract characteristics of human mobility patterns. Recent work has found that individuals are predictable, unique, and slow to explore new places [98, 37, 72, 207, 204, 44, 41]. The availability of similar data nearly anywhere in the world has facilitated comparative studies that show many of these properties hold across the globe despite differences in culture, socioeconomic variables, and geography. The benefits of this data have been realized in various contexts such as daily mobility motifs [191, 198], disease spreading [23, 239] and population movement [141]. While these works have laid an important foundation, there still is a need to integrate these data into transportation planning frameworks.

To make these new data useful for urban planning, we must clarify their biases and build on the progress made by transportation demand modeling even in the face of limited data resources. We must combine this domain knowledge with new algorithms and metrics to better understand travel behaviors and the performance of city infrastructure and we must update technologies to accommodate the computational requirements of processing massive geospatial data sets. Individual survey tracking and stay extraction [10], OD-estimation and validation [39, 164, 230, 118], traffic speed estimation [18, 246], and activity modeling [172, 176] have all been explored using new massive, passively collected data. However, these studies generally present

alternatives for only a few steps in traditional four-step or activity based models for estimating travel demand or fail to compare outputs to travel demand estimates from other sources. Moreover, many methods offered to date lack portability from one city to many with minimal additional data collection or calibration required.

Here we fill this gap with a modular, efficient computational system that performs many aspects of travel demand estimation billions of geo-tagged data points as an input. We review and integrate new and existing algorithms to produce validated origin-destination matrices and road usage patterns. We begin by outlining the system architecture in section 5.2.1. In section 5.2.3 we explain our methods of extracting, cleaning, and storing road network information from a variety of sources. We discuss recent advances in OD creation from mobile phone data in section 5.3.1 and implement a simple, parallel incremental traffic assignment algorithm for these trips in section 5.3.2. We present comparisons of these results to estimates from traditional survey methods in section 5.4.1. Finally, in sections 5.4.2, 5.4.3, 5.4.4 we present a variety of measurements that can be made with the proposed system as well as an online, interactive visualization for conveying these results to researchers, policy makers, and the public. To demonstrate the flexibility of the system, we perform these analyses for five metro regions spanning countries and cultures: Boston and San Francisco, USA, Lisbon and Porto, Portugal, and Rio de Janeiro, Brazil.

5.1.1 Description of Data

Travel surveys are typically administered by state or regional planning organizations and are integrated with public data such as census tracts and the demographic characteristics of their residents, made available by city, state, and federal agencies. New data sources, however, come from new providers. Large telecommunications companies, private applications, and network providers collect and store enormous quantities of data on users of their products and services, presenting computational challenges for storing and analyzing them. Billions of phone calls must be processed, data from open- and crowd- sourced repositories must be parsed, and results must be made more accessible to individuals that generated them. At the same time, it is critical that

measurements from these new sources are statistically representative and corrected for biases inherent in new data. This process requires integration of new pervasive data with reliable (though less extensive) traditional data sources such as the census or travel surveys. We combine the following data sets to illustrate the capabilities of the system architecture here proposed:

1. *Call Detail Records (CDRs):* At least three weeks of call detail records from mobile phone use across each subject city. The data includes the timestamp and the location for every phone call (and in some cases SMS) made by all users of a particular carrier. The spatial granularity of the data varies between cell tower level where calls are mapped to towers and triangulated geographical coordinate pairs where each call has a unique pair of coordinates accurate to within a few hundred meters. Market shares associated with the carriers that provide the data also vary. Personal information is anonymized through the use of hashed identification strings. For reference, 6 weeks of CDR data from the Boston area containing roughly 1 billion calls made by 1.6 million unique users consumes roughly 70 gigabytes of disk space in its raw format. In cities with longer observation periods, data size quickly becomes a performance issue.

2. *Census Data:* At the census tract (or equivalent) scale, we obtain the population and vehicle usage rate of residents in that area. For US cities, the American Community Survey provides this data on the level of census tracts (each containing roughly 5000 people). Census data is obtained for Brazil through IBGE (Instituto Brasileiro de Geografia e Estatística) and for Portugal through the Instituto de Nacional de Estatística. All cities analyzed in this work have varying spatial resolutions of the census information.

3. *Road Networks:* For many cities in the US, detailed road networks are made available by local or state transportation authorities. These GIS shapefiles generally contain road characteristics such as speed limits, road capacities, number of lanes, and classifications. Often, however, these properties are incomplete or missing entirely. Moreover, as such road inventories are expensive to compile

and maintain, they simply do not exist for many cities in the world. In this case, we turn to OpenStreetMaps (OSM), an open source community dedicated to mapping the world through community contributions. For cities where a detailed road network cannot be obtained, we parse OSM files and infer required road characteristics to build realistic and routable networks. At this time, the entirety of the OSM database contains roughly 4 terabytes of geographic features related to roads, buildings, points of interest, and more.

4. *Survey and Model Comparisons:* Wherever possible, we obtain the most recent travel demand model or survey from a particular city and compare the results to those output by our methods. In Boston, we use the 2011 Massachusetts Household Travel Survey (MHTS) and upscale trips according to standard four step model procedures, in San Francisco, the 2000 Bay Area Transportation Survey (BATS), in Rio de Janeiro, a recent transportation model output provided by the local government, and in Lisbon, the most recent estimates from the MIT-Portugal UrbanSim LUT model that uses the 1994 Lisbon transportation survey as input[87]. We found no recent travel survey or model for Porto.

Table 5.1 compiles descriptive statistics for these data sources for each city we explore in the latter sections of this paper.

Table 5.1: A comparison of the extent of the data involved in the analysis of the subject cities.

	City				
	Boston	SF Bay	Rio	Lisbon	Porto
Population (mil.)	4.5	7.15	12.6	2.8	1.7
Area ($1000km^2$)	4.6	18.1	4.5	2.9	2.0
# of Users (mil.)	1.65	0.43	2.19	0.56	0.47
# of Calls (mil.)	905	429	1,045	50	33
# of cell towers	N/A	892	1421	743	335
# of Edges (ths.)	21.8	24.3	22.7	28.1	15.1
# of Nodes (ths.)	9.6	11.3	22.1	16.1	8.6
# of Tracts	732	1139	729	295	272

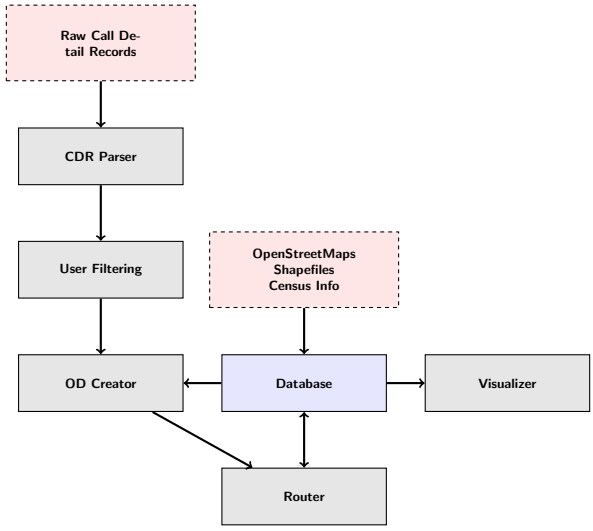

Figure 5-1: A flowchart of the system architecture.

5.2 System Architecture and Implementation

5.2.1 Architecture

The system architecture to integrate the data sources above must be flexible enough to handle different regions of the globe which may have different data availability and quality and efficient enough to analyze massive amounts of data in a reasonable amount of time. The proposed system must also be modular, so that components can be updated easily as new technologies and algorithms become available. To meet these requirements, we choose an object-oriented approach with loose schema requirements. A final object is to make results accessible to a range of end users via online, interactive visualization. To satisfy these constraints, we propose the system architecture depicted in Figure 5-1.

5.2.2 Parsing, Standardizing, and Filtering User Data

One of the biggest challenges in parsing and analyzing travel survey data is the incredible variety in data schema, collection, and reporting practices. Each planning organization typically constructs its own set of data codes and definitions and provides

data in unique formats. This makes it very difficult to compare surveys done in different cities. Call detail records, on the other hand, are typically available for many cities from the same provider and in the same format, and in most cases, translating between the formats of different carriers is simply a matter of shuffling columns. The first component of our system is a simple architecture to convert all CDR data to a standard format that can be expected by the rest of the components.

Given the size of these data sets and the rapidly evolving schema requirements of new models, choosing the proper data structure is critical. Google's open source Protocol Buffer library[1] is an ideal choice as they provide fast serialization for speed and space efficient file storage as well as flexible schemas that can be changed without compromising backwards compatibility. These structures were designed to serve some of the largest databases in the world and are more than enough for our task.

We take a user centric approach to CDR data. We define a *user_data* protocol buffer message that will form the core data structure for our custom User class in an object-oriented programing model. Each User object can be assigned a number of attributes such as the number of calls they make, their home and work locations, and mobility characteristics such as the average time between calls or the average distance traveled on each trip. More sophisticated methods can compute the number and distribution of their trips and even expand them based on census information. We define similar structures and classes for OD matrices, trips, and census data. The serialization routines built into the protocol buffer library ensures that storage of raw data is efficient. To analyze a new city, the user only needs to write two simple routines, one to parse a single line of the CDR file and populate relevant user attributes and one to populate census data objects. Standardizing the CDR data format in this way makes it very easy to compare the output of our estimation models across different cities.

[1]Google Protocol Buffers `https://developers.google.com/protocol-buffers/`

5.2.3 Creating and storing geographic data

A relational database is used to store road network and census information for every city in a standard format. Given the current cost of computing resources, these systems provide adequate performance for storing static GIS and census data and have convenient, mature interfaces for easy access. We also use this database to store aggregated results from our estimates so that they can be made available to interactive web APIs and visualization platforms. We use a Postgres and the open source spatial extension PostGIS to store and manipulate census and road network data.

While census tract or TAZ (Traffic Analysis Zone) polygons and demographic information are stored in this database, it is computationally inefficient to perform point-in-polygon calculations for each user or call record in our CDR dataset. To dramatically speed these computations, we rasterize polygons into a small pixel grid, where pixel values is a unique identifier for the census tract covering that pixel. This raster is then used as a look-up table to convert the latitude and longitude of calls into census tract IDs. The rasterization introduces some error along the boarders of tracts, but these errors are minimized by making pixel sizes much smaller than the size of the raster and resolution of the location estimates of calls (between 10m and 100m).

While the platform supports road networks supplied by local municipalities in the form of shapefiles, we have implemented a parser to construct routable road networks from OpenStreetMap (OSM) data due to its global availability. Transportation networks in OSM are defined by *node* and *way* elements. Nodes represent points in space that can refer to anything from a shop to a road intersection, while ways contain a list of references to nodes that are chained together to form a line. In our context, relevant ways are those used by cars and relevant nodes are intersections within the road network. Ways and nodes may also contain a number of tags to denote attributes such as “number of lanes" or “speed limit". Many roads, however, do not include the whole set of attributes necessary for accurate routing. For example, city roads

often lack speed limit information required to estimate the time cost, which in turn is used to find shortest paths based on total travel time. To infer this missing data, our system supports the creation of user-defined mappings between highway types and road properties. For example, ways tagged as “motorways" are generally major highways and have a speed-limit of 55 mph in the Boston area. They tend to have 3 lanes in each direction. “Residential" roads, on the other hand, have a speed-limit of 25mph and 1 lane in each direction. Each road segment is also given a capacity based on formulas suggested by the US Federal Highway Administration. Using these mappings, we parse the OSM xml data to create a routable, directed road graph with all properties required to estimate realistic costs driving down any given road.

We implement two additional cleaning steps to improve efficiency. The first filters out irrelevant residential roads. These small local roads are filtered from our network, as they are not central to the congestion problem, yet tend to increase computation time significantly. Finally, in OSM data, a node object can refer to many things, for example an actual intersection or simply a vertex on a curve used to draw a turn. The latter case results in a network node with only one incoming and one outgoing edge (assuming U-turns are not allowed). These nodes are superficial and increase network size and routing algorithm run times needlessly. We simplify networks by removing these nodes from the network and only connecting true intersections, keeping the geographic coordinates of the nodes so that link costs still reflect actual geographic length of roads rather than straight line distances between start and end points. The parsed and cleaned edges are then loaded into the Postgres database, preserving attributes and geometry. Pseudo-code of the algorithm to parse and simplify OSM networks can be found in Algorithm 5.1.

5.3 Estimating Origin-Destination Matrices

The following sections review algorithms for transforming billions of geo-tagged data points into validated origin destination matrices and assigning these flows to transportation infrastructure. Some of these algorithms are important for their deviation

from traditional approaches and some are important for their computational efficiency, a requirement when faced with such massive data sets.

5.3.1 Measuring Flow

Current methods to estimate the flow of people or vehicles from place to place in a city generally fall into two categories: four-step or activity based approaches. The former class of models breaks the process into a sequence of four steps from which it earns its name. The first three steps in a four-step model – trip generation, distribution, and mode choice – are designed to estimate origin-destination matrices containing the number of trips from place to place within a city. Traditional modeling approaches use data from travel surveys possibly combined with land use and point of interest information to generate estimates of trip production and attraction for locations. These trips are then distributed from their point of origin to destinations across the city using gravity or radiation models. Modes of transit are assigned using models estimated from survey data and information on the transit infrastructure. More recent activity based models approach travel demand from an individual level. Assuming that travel demand is created by the need to fulfill activities, these models use similar survey data to estimate utility curves for travels and predict behaviors using probit or logit models based on these preferences.

While new data sources such as CDRs do not provide the same detailed demographic and contextual information about individuals or trips, they do provide an opportunity to measure travel more directly. With billions of data points, high spatio, temporal resolution, and long observation periods, passive data collected by mobile devices provide unparalleled scale of observation. New methods to estimate travel demand must balance trade offs between small, but complete data for a short period of time and large, but incomplete data over a longer period of time. In both cases noise and biases must be carefully dealt with to produce valid measurements. In this section we adapt and integrate previous works that have tackled parts of this problem into a full implementation of travel demand estimation for cities.

Initial methods by Wang et al.[230] construct *transient origin-destination matrices*

by simply counting a trip for pair of consecutive calls made within the same hour from two different towers. However, this method lead to an abundance of short trips and provided a very biased view of movement. Instead, mobile phone trajectories must be de-noised to remove spurious points or calls made in the the middle of routes rather than origins or destinations. To extract meaningful locations, termed as *stays*, algorithms have been developed to smooth out this noise and control for these biases. Jiang et al. provide a thorough review of these techniques in [122] and we adapt the stay point algorithm originally described by Zheng et al. in [247].

Given a user's trajectory of spatiotemporal points $P = \{p_1(x_1, y_y, t_1), \ldots, p_n(x_n, y_n, t_n)\}$, the goal is to discover meaningful locations at which a user repeatedly stays for a significant amount of time. The algorithm begins by considering each call in a time ordered sequence. Two consecutive (p_i, p_{i+1}) points are considered to form the start of a *candidate set* of points at the same semantic location if the distance between them is less than a threshold $\Delta r_{i,i+1} < \delta$. Subsequent points are added to this candidate set if they also meet this criteria, e.g. p_{i+2} is added if $\Delta r_{i+1,i+2} < \delta$. The result is a candidate set $S = \{p_s(x_s, y_s, t_s), \ldots, p_t(x_t, y_t, t_t)\}$ containing a number of consecutive calls. A candidate set is considered to represent a single *candidate stay* if time between the first and the last observation in the subsequence S are separated by a time greater than a threshold $\Delta t_{m,n} > \tau$. The geographic location of a candidate stay is set to be at the centroid of points in S. Due to noise in locations and daily call frequencies, multiple candidate stays that are actually the same place may be estimated at a slightly different geographic coordinate on different observation days. To account for this, a final agglomerative clustering algorithm is used to consolidate candidate stays to a single semantic location regardless of the temporal sequence of individual calls. Though many agglomerative clustering algorithms exist, we implement a simple, efficient grid based approach by assigning each filtered location to a grid cell and then defining a final *stay point* as the centroid of all filtered locations in each cell. A final pass through the original calls assigns any call within a distance δ from a stay point to that stay point regardless of whether or a not a consecutive call was recorded from that location. This algorithm removes noisy or spurious out-

liers from the data set while preserving as much information on visits as possible. It may also be run on both triangulated and tower-based CDR data, in the latter case it removes noise associated with calls from the same location being routed through different nearby towers due to environmental factors. Pseudo-code can be found in Algorithm 5.2.

With de-noised trajectories of stay points, the next step is to infer contextual information about each location. Alexander et al. and Colak et al. [7, 47] improve on methods by Wang et al. and Iqbal et al. [230, 118] by using visit frequencies and temporal data to infer contextual information such as a location's function or trip purpose. A user's *home* location is defined as the stay point they are observed at most frequently between the hours of 8pm and 7am on weeknights. Their *work* location is defined as the stay point other than home that a users visits the most between the hours of 7am and 8pm on weekdays. Because many individuals do not work, we leave the work location blank if the candidate location is not visited more than once per week or if the location is less than 500m from their home location. All remaining non- home or work stay points are designated as *other*.

Daily trips are estimated from filtered users by analyzing consecutive observations at different stay points during a given time window. They begin by defining an *effective day* as a period between 3am one morning and 3am on the next consecutive morning. This definition is used to minimize the number of trips that are prematurely ended due to the assumption that users start and end each day at home. A home-based work (HBW) trip is counted if a user is observed to travel between home and work, a non-home based (NHB) trip is counted if a user moves between two non-home stay points, and a home-based other (HBO) trip is counted if a user is observed moving between their home location and a location labeled as other.

Though a user must have traveled between two different observed stay points at some in time, we do not know the precise departure time. We assign a random departure time based on the conditional probability that user departed during an hour between the time they were last observed at the origin and the time they were first observed at the destination. This conditional probability function for departure

time can be derived from surveys such as the National Household Travel Survey or estimated empirically using observed call frequencies of all users over the course of the day. Alexander et al. show that this method produces CDR trip departure time distributions in line with multiple surveys for the Boston region. Having assigned departure times and purposes to each trip, we can construct trips made by a given user. Generally, we are interested in trips between geographic areas such as towns or census tracts so here we convert origin and destination points to IDs of the tract of zone they are in. The result is a vector of trips between locations in the city for each user in our data set.

While a trip represents an observation of movement of at least one person between two locations, we expand these trip counts to represent all individuals in a city. Expansion is a critical step in models relying on survey data where the sample sizes are typically less than 1% of the population. Here we generally have hundreds of thousands of users in our sample, but must still be careful to control for differences in market share and usage rates across a city. We first scale trips based on how often an individual uses their phone. For each user, we calculate the average number of trips made during a given time window by dividing the number of trips counted by the number of days that user was observed making a call. This step effectively measures the average number of trips a user makes between two locations on a day given that they are observed in our data set.

Due to differences in daily usage of mobile phones among the population, not every user makes enough calls on a typical day to infer their movement patterns. For this reason, we must filter out users that do make enough calls. This step requires trade-offs between sample size and amount of data we have on each selected user. Because we will eventually be routing these trips through the transportation network, it is important to correctly estimate the total number of trips taken as well as the distribution of trips across the city. In practice, we find that filtering out users who we measure to make fewer than 2.5 trips per day leaves a large sample size of active users and results in valid estimates of trip tables and OD matrices as shown in subsequent sections. Those implementing these methods may find that different filtering criteria

produce samples suited for different tasks.

We then expand the average trip counts of filtered users to account for market penetration rates. As with survey participants, the ratio of cell phone users to the population is not uniform within the region. Each user is assigned a home census tract and expansion factors are computed for each tract by measuring the ratio of the number of users assigned there and the reported population. In cities such as Boston, these expansion factors tend to be less than 10, but can be higher in places with lower market share. They are generally much lower than surveys which may only choose two or three individuals to represent hundreds or thousands in an area. Each user's typical daily trip volumes are then multiplied by the expansion factor corresponding to their home tract and the now represent the movements of some fraction of the tracts population.

Finally, we may wish to consider only trips via a certain mode, e.g. vehicle trips. Though CDR data does not provide resolution required to measure mode choice, vehicle trips can be approximated by weighting person trips by vehicle usage rates in the home census tract of users. In this way, full OD matrices for vehicle or person trips are computed by summing the expanded trip volume computed for all users between all pairs of census tracts. We also construct partial OD matrices containing only trips of a certain purpose during a certain time window. Due to the relative consistency of CDR data around the world, we can adopt this same OD creation procedure in all cities. Pseudo-code to generate OD matrices has been adapted from [7, 47] and can be found in Algorithm 5.3. The results from this method are compared to the output of traditional models where applicable. Trip tables and correlations plots can be found below in section 5.4.1.

5.3.2 Trip Assignment

Having estimated OD flows, our next task is to efficiently assign these trips to transportation infrastructure, in this case a road network [21]. The first step takes tract to tract OD matrices and distributes trips among nodes, or intersections. Following Wang et al. [230], a trip originating in a census tract is assigned uniformly at random

to an intersection in that tract and to an intersection within its destination tract. This distributes flows such as not to create artificial congestion points and reflects general uncertainty in the exact origin of trips. Other approaches, however, may consist of using abstract centroid nodes unique to each tract and connect to a number of other intersections within that tract using what's referred to as centroid connectors. With intersection to intersection flows, the next task is to assign traffic to routes.

Traffic assignment is another mature domain that has been studied extensively by urban and transportation planners. Static non-equilibrium models approaches consist of treating all users as homogenous agents who make route choices prior to departure based on some heuristic related to current traffic conditions (e.g. the path that minimizes travel time). Incremental Traffic Assignment (ITA) is a variant of these static non-equilibrium assignment models that assigns batches of trips serially and updates costs between increments, as an improvement over the simplest all-or-nothing assignment methods. However, it is known that dynamic equilibrium models are more realistic in assigning trips as outcomes are closer to the Wardrop principles [232], or Nash Equilibria, where drivers seek paths that minimize their travel time and in the final traffic conditions, no driver has an incentive to change their route. To take a step further from static models, Dynamic Traffic Assignment (DTA) [152] models take an iterative and temporally more coherent approach. The addition of these complexities help model traffic flow at finer granularity, enabling road segments to have different conditions within themselves and consequently the representation of phenomena like congestion spill-back, FIFO principle, and others [63].

Our system is modular so that it may implement any number of traffic assignment algorithms. Here, however, we take a simple ITA approach, as it is computationally efficient for many trip pairs in detailed road networks and allows us to keep track of each vehicle as it is routed through the network. We develop a set of tools to perform large scale routing and traffic assignment using parallelization for speedups. First, the parsed and optimized road network is loaded into a graph object. In our implementation, we use the Boost Graph Library for its flexibility and efficiency. We can then compute shortest paths based on a user defined cost (in this case travel

time on road segments). We choose the A* algorithm among the wide range of shortest path algorithms, as it's widely used in routing on geographic networks for its flexibility and efficiency. The A* algorithm implements a *best-first-search* using a specified heuristic function to explore more promising paths first. The euclidian distance between nodes provides an intuitive heuristic that ensures optimal solutions are found. While this algorithm provides the same results as Dijkstra's algorithm, we find that it becomes more efficient to compute paths one by one for sparse OD matrices.

On most city roads, free-flow speeds are rarely achieved due to congestion. As a result, traffic patterns may significantly change the time costs associated with using a particular route. In addressing this, the most simple procedure is the Incremental Traffic Assignment algorithm [169]. A simplified schematic explaining the procedure can be seen in Figure 5-2. This algorithm assigns trips in a series of increments and updates the costs of edges in the network based on the number of vehicles that were previously assigned to that road between increments. For example, the first increment assigns 40% of trips for each pair assuming each driver experiences free-flow speeds. The travel time cost associated with every road segment is then adjusted based on how many drivers were assigned to that road and the total number of cars a road can accommodate in unit time. The next 30% of drivers are then routed in the updated conditions. This process is repeated until all users have been assigned a route. The shortcoming of this method is that once a driver has been assigned a route it does not change, and consequently the approach does not converge to Wardrop's equilibrium

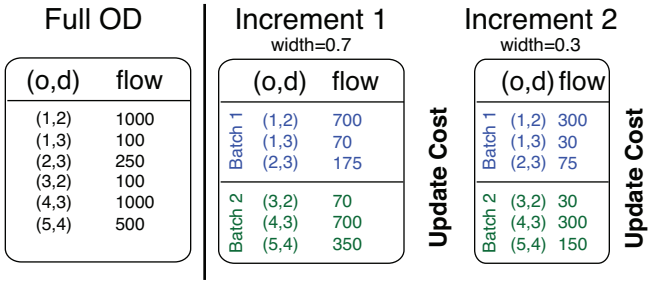

Figure 5-2: Our efficient implementation of the incremental traffic assignment (ITA) model. A sample OD matrix is divided into two increments and then split into two independent batches each.

even for very small increment sizes. Yet we use it here due ease of implementation and the fact that it is still insightful for the purposes of demonstrating the implementation of a modular data-driven four step model. Future work will explore the use of newer methods.

Relating travel performance to traffic conditions has been a long standing problem in transportation. Many different characterizations exist, ranging from conical volume-delay functions to more complex approaches [34, 212, 6]. One of the most simplistic and common metrics used in determining the travel time associated with a specific flow level is the ratio between the number of cars actually using a road (volume) and it's maximum flow capacity (volume-over-capacity or V/C). At low V/C, drivers enjoy large spaces between cars and can safely travel at free-flow speeds. As roads become congested and V/C increases, drivers are forced to slow down to insure they have adequate time to react. Based on the volume-over-capacity (V/C) for each road, costs are updated according to Eq. 5.1, where $\alpha = 0.15,\ \beta = 4$ are used per guidelines set by the Bureau of Public Roads[2].

$$t_{current} = t_{freeflow} \cdot (1 + \alpha(V/C)^{\beta}) \tag{5.1}$$

Though increments must be routed in serial, all routes discovered within an increment are independent. To speed up the routing process, we divide all trips in an increment into batches and send these batches to different threads for parallel computation. Because the road network remains fixed in each increment, we only need to store a single graph object shared by all threads. When a shortest path is found, we walk that path and increment counts of the number of vehicles that were assigned to each road and sum the counts from all batches after the increment has finished. We also keep track of the origin and destination census tracts of the assigned vehicles in a bipartite graph for later analysis. After all trips have been routed, we compute final V/C ratios and other metrics of each segment and update these values in the database so they can be used for other applications or visualization. Pseudo code for

[2]Travel Demand Modeling with TransCAD 5.0, User's Guide (Caliper., 2008).

this ITA procedure can be found in Algorithm 5.4.

5.4 Results

In the following sections we demonstrate the range of outputs provided by our system. We first report trip tables and compare origin-destination matrices produced by our system to available estimates made using travel surveys. We then report road network performance as well as characteristics of road usage patterns enabled by the construction of a bipartite road usage network.

5.4.1 Trip Tables and Survey Comparison

In order to understand when and where these new data will be effective and how the results differ from traditional approaches, we compare the output of our system to previous travel surveys wherever possible. In four of the cities studied, we find estimates of travel demand from surveys: the 2011 Massachusetts Household Travel Survey (MHTS) in Boston, the 2000 Bay Area Travel Survey (BATS) in San Francisco, a 2013 transportation plan in Rio de Janeiro, and estimates from a 2012 LUT model in Lisbon[87]. While these surveys do not always produce all estimates we are able to generate with our system, we make comparisons wherever possible.

Trip tables report the total number of trips of a given purpose or during a given time of day for a city and represent the total load placed on transportation infrastructure. In Table 5.2, we report trip tables for each city in this study. We find close agreement with trip tables estimated using CDR data and surveys in Boston and the San Francisco Bay Area and less agreement in Rio de Janeiro. We note, however, that the 3.74 million person trips estimated for Rio is far too low given the population of the region and highlights the difficulty in finding reliable planning resources in many areas. Finally, we note that in Lisbon, the survey results represent vehicle trips only, while we report person trips. When adjusting for mode car ownership rates in Portugal, our numbers align more closely. We were unable to find a survey or model for comparison in Porto.

Table 5.2: Trip tables estimates. Where possible, our results are compared to estimates made using travel surveys. For each city, we report the number of person trips in millions for a given purpose or time. Trip purposes include: home-based word (HBW), home-based other (HBO), and non-home-based (NHB). Trip periods include: 7am-10am (AM), 10am-4pm(MD), 4pm-7pm (PM), and the rest of the day (RD). We note that the exact boundaries of the surveys do not exactly coincide with those used in our estimation so direct comparisons are not exact. In general, trip magnitudes align closely, with the exception of Rio de Janeiro, where the survey results report far too few trips, illustrating the difficulty of obtaining sensible measurements via certain techniques. No comparisons could be found for Porto.

City	HBW	HBO	NHB	AM	MD	PM	RD	Total
Boston	5.76	8.99	6.72	3.71	7.68	5.75	4.33	21.47
MHTS	3.22	12.83	9.49	5.32	8.87	8.20	3.15	25.54
SF Bay	4.07	10.05	7.04	4.47	7.81	5.35	3.53	21.16
BATS	4.60	11.54	4.66	4.18	6.90	4.22	3.00	20.80
Rio	9.92	17.17	11.46	7.71	14.09	10.47	6.29	38.55
Survey	2.06	–	–	1.31	1.19	1.24	–	3.74
Lisbon	1.08	2.01	1.21	0.79	1.67	1.26	0.58	4.30
Survey[3]	0.61	–	–	–	–	–	–	–
Porto	0.49	0.87	0.46	0.32	0.70	0.54	0.27	1.83
Survey	–	–	–	–	–	–	–	–

In addition to trip tables, it is also necessary to compare the distribution of trips from place to place around the city. In order to make this comparison, the area unit of analysis for the survey and our model must be aligned. Given the resolution of mobile phone data, our system is designed to create ODs at the census tract (or equivalent) level while many surveys aggregate to larger traffic analysis zones or super districts. For comparison, we aggregate the OD matrices from CDRs to the coarser grained resolution provided by the survey and compare results. Figure 5-3 show correlation histograms comparing OD matrices at the largest spatial aggregation available produced by our methods and those produced by traditional methods. In general we find very high correlations in Boston, San Francisco, and Rio, with lower correlations in Lisbon. Lisbon, however, has the smallest units of aggregation and these results demonstrate the limitations of these comparisons at very high spatial resolutions. We hope future work explores how these correlations relate to the modifiable area unit problem. Finally, there is significant uncertainty in all models and we hope future works will explore this uncertainty further.

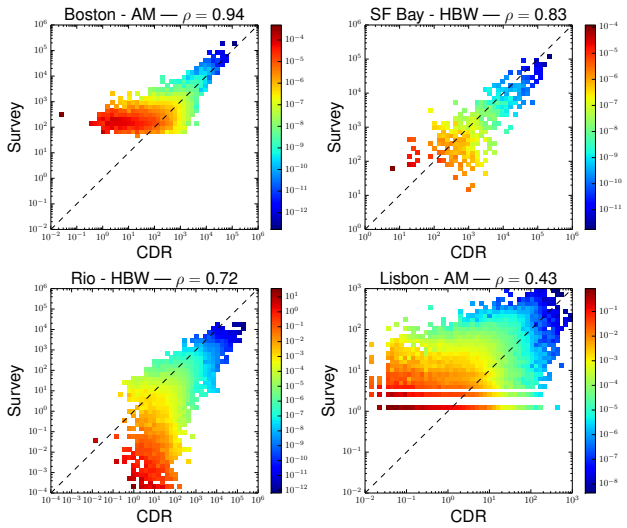

Figure 5-3: Correlations between OD matrices produced by our system and those derived from travel surveys at the largest spatial aggregation of the two models. In Boston, this is town-to-town, in San Francisco, MTC superdistrict-to-super district, in Rio, census superdistrict-to-superdistrict, and in Lisbon, freguesia-to-freguesia. The larger of these area units (e.g. towns in Boston), the better our correlations, while correlations at the smallest aggregates(e.g. freguesias in Portugal), correlations are lower. However, more work must be done to understand uncertainties in estimates provided by both models.

5.4.2 Road Network Analysis

The first output of this procedure is volume, congestion (volume-over-capacity), and travel times for all road segments. Using the outcomes of our analyses, we calculated the distributions of volumes on roads, along with V/Cs in Figure 5-4. Interestingly, the results suggest qualitatively similarly distributed volumes and V/Cs for our five subject cities. Moreover, our findings are consistent with general congestion studies that identify Rio de Janeiro as one of the most congested cities in the world and the San Francisco Bay Area not far behind. Smaller cities such as Boston and Porto have fewer problems with congestion.

5.4.3 Bipartite Road Usage Graph

In addition to measuring physical network properties of roads, the system architecture enables detailed analysis of individual road segments and neighborhoods within a city. Following Wang et al. [230], we create a bi-partite usage graph. Every time a route

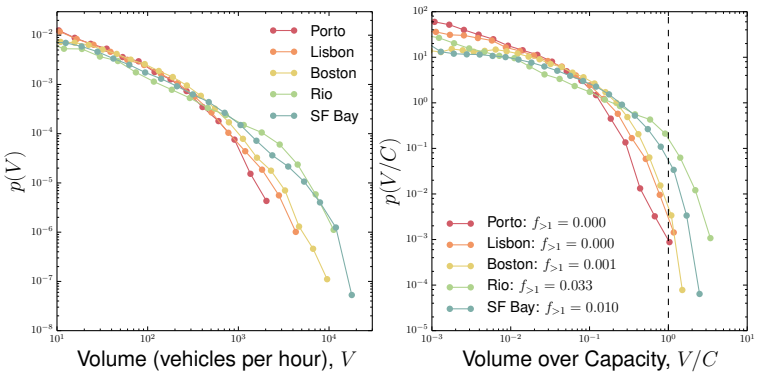

Figure 5-4: Distributions of travel volume assigned to a road and the volume-over-capacity (V/C) ratio for the five cities. The values presented in the legend refers to the fraction of road segments with $V/C > 1$.

between two location is assigned, we traverse the path and keep a record of how many trips from each driver source (census tract) used each road. This record is then used to construct a bipartite graph containing two types of nodes: road segments and driver sources, as shown in Figure 5-5. Roads are connected to driver sources that contribute traffic to that segment and census tracts are connected to roads that are used by people who live here.

$$k_s^{road} = \sum_o A_{o \to s}, \; k_o^{source} = \sum_s A_{o \to s} \tag{5.2}$$

$$A_{o \to s} = \begin{cases} 1, & \text{if vehicles from tract } o \text{ use road } s \\ 0, & \text{otherwise.} \end{cases}$$

We then examine the degree distributions of roads and census tracts using Eq.5.2 in this bipartite graph to reveal patterns of road usage in Figure 5-7. The number of roads used by residents of a given location is much more consistent between different cities and appears less affected by the size of the road network. On the other hand, the number of driver sources contributing traffic to a given road segment is broadly distributed, suggesting that most roads are *local* in that they serve only a few locations, while a few roads in the tail of the distribution are used for large fractions of the population. While this result is intuitive given that highways are designed for just this purpose, we hope future work explores the relationship between this bipartite

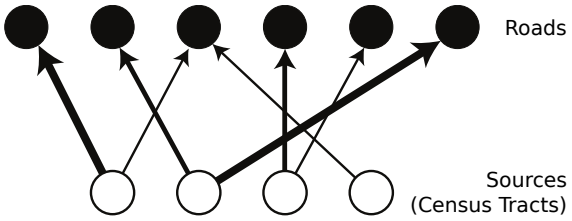

Figure 5-5: A graphical representation of the bipartite network of roads and sources (census tracts), with edge sizes mapping the number of users using the connected road in their individual routes.

usage graph and road network topology further.

An example of such an application was proposed by Wang et al. to classify road segments based on the relationship between topological and demand based metrics. Comparing the topological properties of roads in the physical network to the bipartite usage graph provides insights into their role in the transportation system. Edge betweenness centrality [162] captures the importance of a road by counting how many shortest paths between any two locations σ_{OD} must pass through that edge $\sigma_{OD}(e)$ (Eq. 5.3). While this measure captures some aspects of importance, it treats all potential paths as equally likely and tends to be biased towards geographically central links. The degree of a road in the bipartite usage graph reflects the number of locations in the city that actually rely on that road because trips were assigned there from actual travel demand. With these two metrics, betweenness centrality and a roads degree in the usage network, we can classify the role of a road in the cities transportation network.

$$bc_s = \sum_{o,d} \frac{\sigma_{OD}(s)}{\sigma_{OD}} \tag{5.3}$$

A simple classifier divides the betweenness usage degree space into four quadrants surrounding the point representing the 75th percentile for betweenness centrality and usage degree. Roads with betweenness and usage degree above the 75th percentile are both physical connectors and are used by large portions of the region. These roads tend to be bridges or urban rings. Roads with low betweenness, but high usage degree are attractors, receiving a higher proportion of trips than would be expected

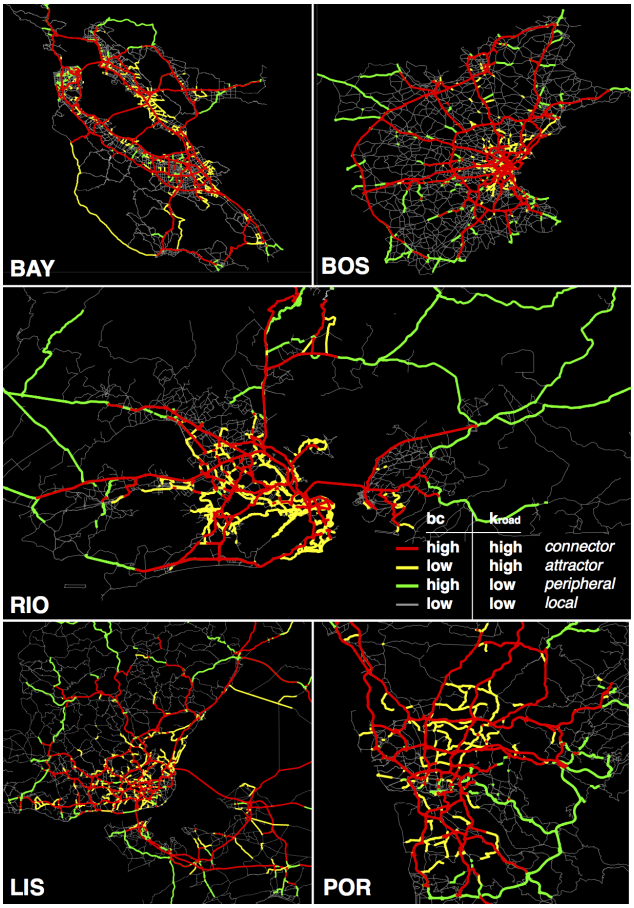

Figure 5-6: Maps depicting the proposed road classification, summarized in the legend, for the five subject cities.

assuming uniform demand. Roads with high betweenness and low usage are physical connectors and serve an important purpose geographically, but may not be utilized by actual demand. Other roads, with low betweenness and low usage are local roads and primarily serve populations living and working nearby. Figure 5-6 shows each road according to this classification using data from the ODs calculated via mobile phones.

Finally, this bipartite framework of analysis allows us to augment visualizations of congestion maps in two ways. The first focuses on a single road segment. For example, when we identify a segment of a highway that becomes highly congested with traffic jams each day, we can easily query the bipartite graph to obtain a list of census tracts where drivers sitting in that traffic jam are coming from and where they are going to. The census tract nodes can also be given attributes from containing

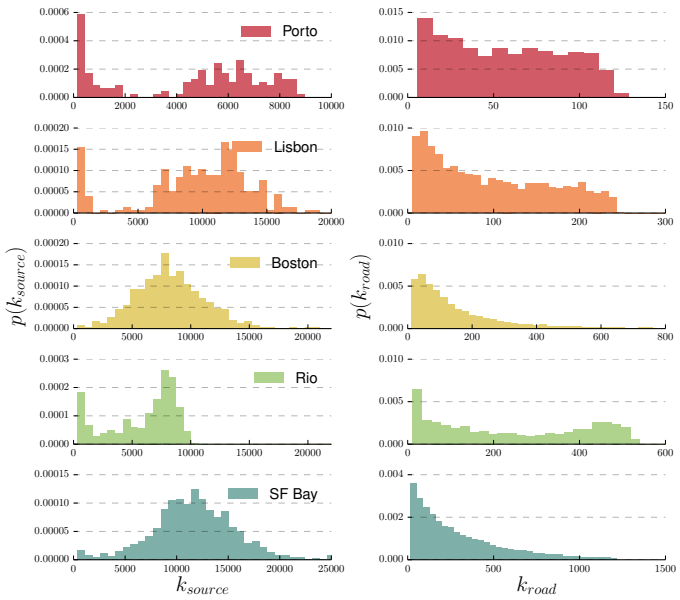

Figure 5-7: Distributions of k_{road} and k_{source} for the five cities. Inset: The unitized collapsed normal distribution for k_{source}.

Figure 5-8: Two screen images from the visualization platform. (a) The trip producing (red) and trip attracting (blue) census tracts using Cambridge St., crossing the Charles River in Boston. (b) Roads used by trips generated at the census tract including MIT.

any demographic data a user wishes. With this information, it is possible to identify leverage points where policy makers can offer alternatives to these individuals or even power applications such as car sharing, by notifying drivers that others sharing

the same road may be going to and from the same places. Moreover, businesses considering products or services based on who may be driving by or near different locations may find value in these detailed breakdowns.

Rather than selecting a road segment node, we may also select a single census tract, and check its neighbors to construct a list of all roads used by individuals moving to or from that location. For example, for a given neighborhood in a city we can identify all major arteries that serve that local population. This information provides a detailed look at a central location based on how much road usage it induces. Moreover, geographic accessibility, critical to many socio-economic outcomes, can now be measured in locations that were previously understudied.

5.4.4 Visualization

To help make these results accessible to consumers and policymakers, we build an interactive web visualization to explore road usage patterns in each city. Most GIS platforms can connect directly PostGIS databases to visualize and analyze road networks with our estimated usage characteristics. While these platforms are preferred by advanced users familiar with GIS data, they are opaque to many consumers who may benefit from more detailed information on road usage. A simple API is implemented to query the database and generate standard GeoJSON objects containing geographic information on roads as well as computed metrics such as level of service. We also implement queries to answer questions such as “What are all the census tracts used by drivers on a particular road?" or "What are all roads used by a given location in the city?". These data are then parsed and displayed on interactive maps using any of the available online mapping APIs and D3js allowing users, with functionality that enables one to select individual roads and areas. Two screen images of this system is shown in Figure 5-8.

5.5 Limitations and Future Work

This chapter has presented a full implantation of a travel demand model that uses new, big data resources as input. We have presented a system that combines and improved upon many disparate advanced in recent years to produce fast, accurate, and inexpensive travel demand estimates. We began by outlining methods to extract meaningful locations from noisy call detail records and estimate origin-destination matrices by counting trips between these places. Normalized and scaled trips counts are compared to estimates made using survey data in both trip tables and at the OD pair level. These flows are then assigned to road networks constructed from OpenStreetMap data using an incremental traffic assignment algorithm. As routes are assigned, a number of metrics on road usage are measured and stored.

While these results show great progress in making big data useful for transportation engineering, there are still limitations inherent in this data and our model. Specifically, we highlight three areas that are ripe for further study.

1. We have shown the the level of aggregation applied to OD matrices can affect the correlation observed between model outputs. This is a standard manifestation of the modifiable area unit problem and a more detailed exploration may indicate which levels of analyses were better suited for different data sources. Moreover, a more detailed analysis of uncertainty in model estimates may make it easier to assess their correlation and validity.

2. Our traffic assignment algorithm is efficient, but simple. In the future, a stochastic dynamic user equilibrium assignment methods should be explored and compared. Moreover, route choice modeling may be significantly improved by the availability of high resolution GPS trajectories of drivers. We believe our system's modular design makes it easy to incorporate these new models.

3. Our mode choice model remains simple and will likely require more sophistication for modeling trips not taken in private vehicles. This, combined with improvements in route choice, may make it possible to estimate multi-modal

trip demand, as public transportation, bike lanes, and even water transportation networks are included in OpenStreetMap data.

We hope future work will address these and improve further on the methods presented here.

5.6 Conclusions

Transportation engineers and urban planners have a rich history estimating flows of people within cities and mapping this flow onto transportation infrastructure. However, these efforts are often constrained by limited data resources. The rise of ubiquitous mobile sensors has generated a wealth of new data on human mobility, but new tools must be developed to integrate these data and insights into traditional transportation modeling approaches. To this end, we have demonstrated a full implementation of a travel demand model utilizing mobile phone data as an input. We presented algorithms to generate routable road networks from open source data repositories, generate validated OD matrices and trip tables from CDR data, and route these trips through road networks using a paralleled ITA algorithm. We have demonstrated a number of possible analyses that can be performed on the output of this system including network performance and classification measurements and an online, interactive visualization platform.

As more data becomes available in the form of calls, gps traces, or real time traffic monitoring systems, we are excited at the prospect of updating and improving these systems further.

5.7 Acknowledgments

The work in this chapter was the result of a collaboration with Serdar Colak, Bradley Sturt, Lauren P. Alexander, Alexandre Evsukoff, and Marta C. González. It was partially funded by the BMW-MIT collaboration under the supervision of PI Mark

Leach[4], the World Bank-HuMNet collaboration agreement under the supervision of PI Shomik Mehndiratta[5] and the Center for Complex Engineering Systems (CCES) at KACST under the co-direction of Anas Alfaris[6]. Many thanks to Pu Wang for technical support, Shan Jiang for her help obtaining LUT model results for Lisbon, Nelson F. F. Ebecken for support with data, the Rio de Janeiro State Agency (FAPERJ) for the grant on this project, and the Rio City Hall for the support and the data they have provided.

5.8 Chapter 5 - Appendix

5.8.1 Algorithms

ALGORITHM 5.1: Parsing OpenStreetMap Networks

```
{OSM files are XML based and contain way and node objects}
ways = set of ways in an OSM file
nodes = set of nodes in an OSM file
graph = an empty graph

{Add each pair of consecutive nodes to the edge list}
for way in ways do
    for i = 0 to i = way.nodes.size() − 2 do
        graph.addNode(way.nodes[i])
        graph.addNode(way.nodes[i + 1])
        graph.addEdge(way.nodes[i], way.nodes[i + 1])

{Simplify the network by merging road segments }
for way in ways do
    startNode = way.nodes[0]
    for node in way.nodes do
        if all edges into and out of node are segments of the same way then
            graph.removeNode(node)
            remove all edges to or from node
        else
            endNode = node
            graph.addEdge(startNode, endNode)
            FillEdgeAttributes()
            startNode = endNode

{Notes}
*FillEdgeAttributes() fills in missing data such as speed limits or number of lanes based on way attributes
*graph.addNode(node) and graph.addEdge(node1, node2) only add objects if they do not already exist
*graph.removeNode(node) also removes all edges containing that node
*when simplifying the network, proper geographic lengths are kept even when nodes are deleted
```

[4]mark.leach@bmw.de

[5]smehndiratta@worldbank.org

[6]anas@mit.edu

ALGORITHM 5.2: Stay Point Algorithm

```
{Each user object has a number of attributes}
call = a call object with an associated latitude, longitude, stay index
calls = vector of a user's calls ordered by timestamp
candidateSet = empty set of consecutive calls that meet criteria for a stay
candidateStays = a vector of centroids from candidate sets
δ = distance threshold between consecutive calls (in meters)
τ = time threshold between entry into and exit from the stay (in seconds)
ds = a grid size for the agglomerative clustering algorithm (in meters)
stayCalls = an empty vector of calls from stay points
---------------------------------------------------------------------------
{For each user, loop through all calls and find candidate stays}
candidateIndex = 0
candidateSet = {}
for i = 0 to i = calls.size() − 2 do
    if DistanceBetweenCalls(calls[i], calls[i + 1]) < δ then
        candidateSet.append(calls[i + 1])
    else
        if TimeBetweenCalls(candidateSet[0], candidateSet[end]) > τ then
            for call in candidateSet do
                call.stayIndex = candidateIndex
            candidateStay = Centroid(candidateSet)
            candidateStays.append(candidateStay)
        candidateSet = {calls[i]}
        candidateIndex = candidateIndex + 1
---------------------------------------------------------------------------
{Run an agglomerative clustering algorithm}
grid = construct a uniform grid that covers all of a user's calls with cell dimensions ds × ds
stayIndex = 0
for grid cells containing a candidateStay do
    candidateStays = {listof candidateStay incell}
    stay = Centroid(candidateStays)
    for call made from a candidateStay in this cell do
        call.longitude = stay.longitude
        call.latitude = stay.latitude
        call.stayIndex = stayIndex
        stayCalls.append(call)
    stayIndex = stayIndex + 1
---------------------------------------------------------------------------
{Final pass to add any remaining calls to the stay}
for i = 0 to i = calls.size() do
    if call not part of a stay and DistanceBetweenCalls(call, stay) < δ for any stay then
        call.longitude = stay.longitude
        call.latitude = stay.latitude
        call.stayIndex = stayIndex
        stayCalls.append(call)
Sort stayCalls by timestamp
---------------------------------------------------------------------------
{Notes}
*Centroid(callSet) returns an object whose latitude and longitude are the centroid of all points in the input
*DistanceBetweenCalls(call1, call2) returns the geographic distance between calls in meters
*TimeBetweenCalls(call1, call2) returns the time between call in seconds
```

ALGORITHM 5.3: OD Creation Algorithm

```
{Data objects}
tracts = census tract data objects containing demographic variables
OD(o, d, p, t) = 0 for origin o, destination d, purpose p, and period t
_____
{Detect home and work for all users and compute expansion factors}
for user in users do
    user.stays = vector of calls at stay points sorted by time
    user.home = index of stay point visited the most between 8pm and 7am on weekdays
    user.work = index of non-home stay point visited the most between 7am and 8pm on weekdays
    if user visits work less than once per week then
        user.work = null
    for stay in user.stays do
        stay.label assigned as home, work, or other
    user.weekdays = number of weekdays a user records a stay
    user.workdays = number of weekdays a user records a stay at work
    tract[user.home].numUsers = tract[user.home].numUsers + 1
for tract in tracts do
    tract.expansionFactor = tract.population/tract.numUsers
_____
{Count and expand trips}
for user in users do
    trips = empty vector to store trips taken by a user
    for i = 1 to i = user.stays.size() do
        s0 = user.stays[i − 1]
        s1 = user.stays[i]
        if s0 == S1 then
            continue
        if  s0 and s1 are on the same effective day then
            trip = new trip from s0 to s1
            trip.purpose = PurposeFromLabels(s0, s1)
            trip.workday = true if workday for user, false otherwise
            trip.departure = GetConditionalDepartureTime(s0, s1)
            trips.append(trip)
        elses0 and s1 are not on the same effective day
            morning = create trip from home to first recorded stay
            night = create trip from last recorded stay to home
            trips.append(morning)
            trips.append(night)
    for trip in trips do
        o = trip.origin
        d = trip.destination
        p = trip.purpose
        t = trip.departure
        if trip.workday == true then
            flow = tract[user.home].expansionFactor/user.workdays
        else
            flow = tract[user.home].expansionFactor/user.weekdays
        OD(o, d, p, t) = OD(o, d, p, t) + flow
_____
{Notes}
*PurposeFromLabels(s0, s1) returns a trip purpose (HBW, NHB, HBO) based on the label of origin and destination stays
*GetConditionalDepartureTime(s0, s1) returns a departure time based on the observation times at origin and destination
*an effective day is defined as a period between 3am today until 3am on the next consecutive morning
```

ALGORITHM 5.4: Incremental Traffic Assignment

$graph$ = road network
$OD(p,t)$ = origin-destination matrix for purpose p and time window t
B = a bipartite network containing roads and census tracts
$incrSize$ = vector of increment sizes, e.g. [0.4, 0.3, 0.2, 0.1]
$nBatches$ = number of threads to use
for $i = 0$ **to** $i < incrSize.size()$ **do**
    for $b = 0$ **to** $b < nBatches$ **do**
        create new thread
        $batch = GetBatch(OD, b)$
        for all o, d pairs in $batch$ **do**
            $flow = OD[o,d].flow \cdot incrSize[i]$
            $route = A^*(o,d,graph)$
            for all segment s in $route$ **do**
                $s.flow = s.flow + flow$
                $B_{e \to o} = B_{s \to o} + flow$
    wait for all threads to finish
    for segment s in $graph$ **do**
        $s.cost \leftarrow s.freeFlowTime \cdot (1 + \alpha(\frac{s.volume}{s.capacity})^{\beta})$

* GetBatch(OD, B) returns only the subset of OD pairs pertaining to a batch
* $A^*(o,d,graph)$ returns the shortest path between o and d if a path exists

Chapter 6

Inferring land use from mobile phone activity and points of interest

6.1 Introduction

In describing the “organized complexity" of cities, Jane Jacobs notes that a "park's use depends, in turn, on who is around to use the park and when, and this in turn depends on uses of the city outside the park itself." [120] Where people live, work, and play is intimately related to the time and distance required to move to and from places [90]. Understanding how individuals are distributed in space and time is crucial to making effective and efficient planning decisions within cities. The locations of public facilities and private businesses are influenced by and determine the demand for mobility.

How a particular area of a city is used is decided, in part, by the zoning regulations implemented and enforced through local governments. These regulations impact the structure of a city, dictating where housing or office space can be located. Zones of a kind share common usage. The central business district (CBD), for instance, is populated during office work hours whereas when offices are closed, relatively few people are found in these zones. Thus land use is closely is related to differences in population presence to be found at any given time in the zone. In practice, however, many zones feature a variety of uses which may differ from the official designation.

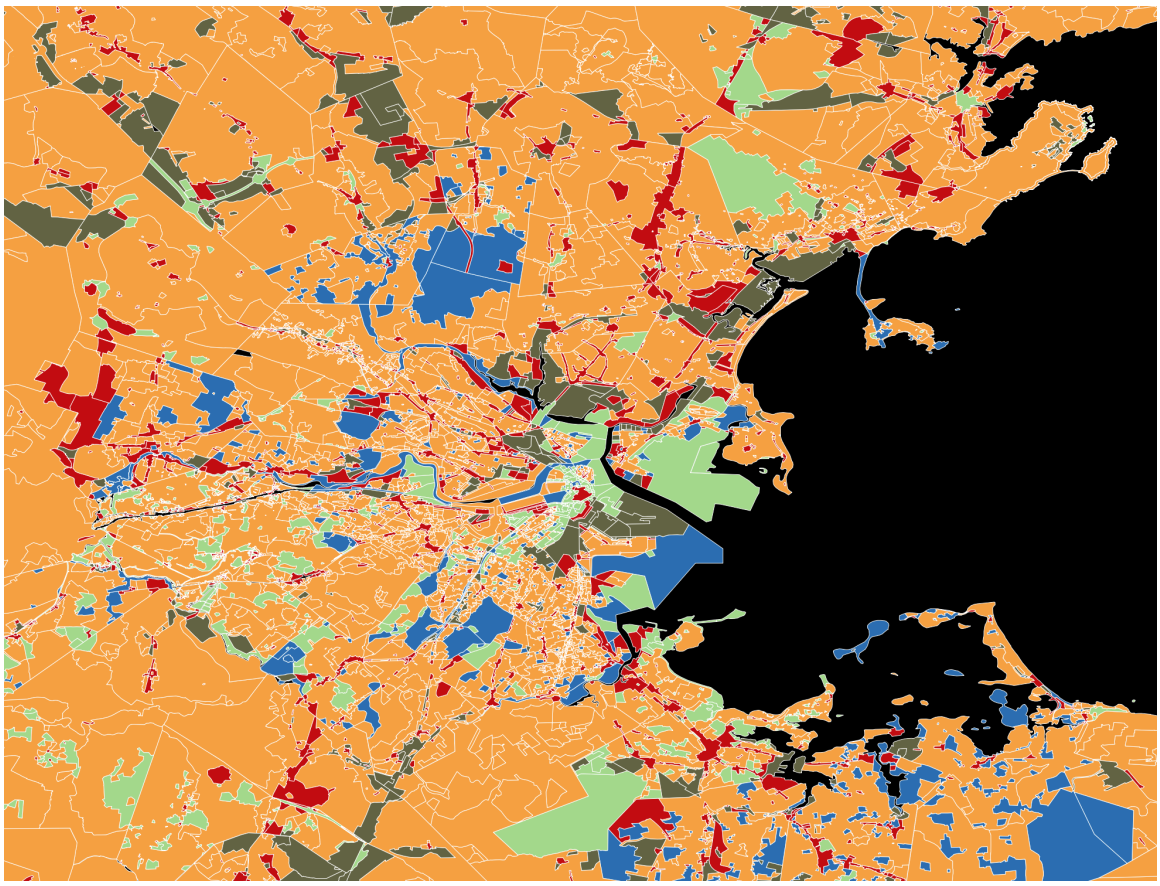

Figure 6-1: Zoning regulation for the Boston area. Color code: orange - Residential, red - Commercial, gray - Industrial, blue - Parks, green - Other.

As an example, zoning information for the Boston area is shown in Figure 6-1. Note, that zoning areas are not only restricted to land but also cover parts of rivers, lakes and the sea.

There is a large body of work dedicated to understanding the spatiotemporal dynamics of population and its relation to land use [143], [17], [49]. Measurements of human mobility within cities have traditionally been made via travel surveys. These surveys require subjects to record where they move to and from over an observation period (typically one day or a whole week), how they do so, and why. However, because surveys typically feature in-person interviews and demand a high workloads from each subject, this method of data collection is expensive and limited.

Given these limitations, travel surveys suffer from relatively small samples (usually below tens of thousands of individuals), capture only short periods for each individ-

ual, and are updated infrequently. Fortunately, over the past decade a new type of measurement instrument has made its way into the pockets of people in nearly every culture and country. Each of the roughly 6 billion mobile phones currently in use [1] is capable of recording the location of calls, SMS, and data transmissions to within a few hundred meters. Moreover, these data are also collected and stored centrally by mobile phone providers for billing purposes. With these data come enormous opportunities to improve our understanding of human mobility patterns.

In particular, call detail records (CDR), which provide information on the location of mobile phones any time a call is made or a text message is sent, provide large amounts information on the distribution of persons in a region at low costs compared to surveys. With such detailed information on the movement and communication patterns of individuals, privacy is an immediate concern. For the purpose of this stud, we aggregate data to measure only the number of active phones in a given area during a given time interval. This method of data collection provides much higher levels of anonymity reduces the risk that any breach of individual information while still revealing insights into the city.

In parallel with the growth in mobile phone use, services such as Google Places and Yelp allow anyone to log and access information on local points of interest (POIs) detailing the exact location and type of businesses, parks, and other important locations across a metropolitan area. These databases provide low cost, current, and detailed information on the places individuals go and what they do when they get there. Given these new data sources, the question arises as to whether the distribution of the numbers of active mobile phones and POIs can be used to quantify the how activity varies with zoning.

To have such a measurement method would be very advantageous. Zoning regulations exist, in part, to control the structure and function of cities. New data sources may provide a method of testing if the actual activity patterns of locations match the desired affect of zoning. In the event that activity patterns are not uniform across similarly designated zones, quantifying difference in usage may help better under-

[1] `http://www.itu.int/net/ITU-D/index.aspx`

stand demand for mobility infrastructure across space and time compared to current analysis. Monitoring the usage over time allows to detect changes in habits of the population as well as shifts in usage which may indicate ongoing regional developments.

Consequently this chapter investigates the potential of aggregated CDR data and point of interest data to infer dynamic land use, i.e. to understand how the population of different areas of a city changes with time and and the type of businesses that exist there according to specific zoned land uses. The chapter centers on supervised classification of regions according to given zoning regulations. We demonstrate that CDR data can be used in order to classify zones of different types with reasonable accuracy. We explore the limitations of this accuracy in comparison to alternative measurements of land use. Through this process, normalization techniques are discussed to highlight differences between zones. The application and result of random forests for the classification is described in detail.

6.2 Mobile phones and human mobility

Mobile phones have proven exceptionally good instruments to measure human behavior with. In one of the first studies utilizing these devices, Eagle and Pentland [80] were able to decompose mobile phone activity patterns of university students and employees into regular daily routines. Moreover, these patterns were found to be predictive of an individual's characteristics such as their major or employment level (i.e. graduate student). Subsequent research has built upon this work, scaling up in both geographic extent and sample size. González et al [97] studied data from nearly one-hundred thousand anonymous mobile phone users to reveal persistent regularities in the statistical properties of human mobility. Highlighting the remarkable predictability of human behavior, Song et al [208] estimated that it is theoretically possible to predict individual movements of users with as high as 93% accuracy using only data from mobile phones.

Mobile phone data has also provided insights on how space is used over time. For

example, Reades et al [175] link mobile phone activity to commercial land uses in Rome, Italy. Measuring activity in 1km by 1km grid cells, they employ a form of principal component analysis to identify the dominant activity patterns. The authors qualitatively interpret areas of the city exhibiting this signal as commercial, though actual zoning information is not introduced. They then decompose activity across the city to identify regions with similar patterns of usage. Soto et al [209] use CDR mobile phone data at the cell tower level to identify clusters of locations with similar activity. Qualitative agreement between these clusters and land uses were observed.

Calebrese et al [43] apply similar decomposition and clustering techniques to classify locations on a university campus as classrooms, dormitories, etc. By analyzing wifi activity across 3000 wifi-access points, the authors used unsupervised, non-parametric techniques to identify clusters of locations with similar internet use. These locations naturally fit into location profiles such as "lecture hall" or "dormitory." Finally, CDR data have proven useful to detect movement at the census tract scale [40]. Location data from calls helped to measure origins and destinations for trips across the Boston Metropolitan area. However, no attempt was made to associate such trips with land uses.

Other data sources such as points of interest (POIs) as well as GPS data collected from taxi fleets have been combined with unsupervised learning algorithms to identify the rich structure of different functional sections of a Beijing [244]. To date, however no studies exist that employ supervised learning techniques to combine traditional data sources on land use such as zoning regulations and CDR data or POI data. This study aims to investigate the link between zoned land use and mobile phone activity on a common spatial partitioning of the greater Boston area into regions of homogeneous land use. For each region the temporal profile of active phones is used in supervised classification techniques in order to identify patterns characteristic for a specific zoning classification. The corresponding patterns will be interpreted in detail.

6.3 Data sources

Three data sources are used in this chapter: mobile phone activity records, digital point of interest logs, and official zoning regulations. For the Boston metro region, anonymized CDRs provide the location of a mobile phone by triangulating signal strengths from surrounding cell towers. Note this differs slightly from traditional CDR data in which record the location of a call as the location of the mobile phone tower. This provides slightly higher accuracy and allows us to measure calls continuously across space rather than at points where towers are located. Triangulation by this method is accurate to within a few hundred meters depending on the tower density. These data make it possible to measure the amount of phone activity (counts of the number of calls and texts) that occurs within a given area and time window. In this study we use three weeks of CDR data for roughly 600,000 users in the Boston region home to roughly 3 million people. Though mobile phone data come from specific set of carriers, the market share of these carriers is between 30% and 50%.

Points of interest (POIs) for the greater Boston region were crawled from the Google Places API. This web application allows businesses to place information such as address, descriptions, hours, etc. on into Google's database. When users search a spatial region for points of interest matching keywords such as "restaurant" or "post office," they are given a list of locations matching that criteria in their area. Each point of interest consists of a location (latitude and longitude) as well as a number of tags describing the business or place (note a location may have more than one tag). The vast majority of POIs are establishments such as businesses or government facilities. Other semantic places such as markers for political zones or important transportation junctions also exist. For the Boston region, our dataset contains roughly 180,000 POIs with 104 unique descriptive tags.

In addition to novel data sources such as mobile phones and online POI databases, we obtain zoning classifications for the Boston metropolitan area. The Massachusetts Office of Geographic Information (MassGIS) aggregates uses into five categories: Residential, Commercial, Industrial, Parks, and Other. We are careful to note our as-

sumption that actual land use and zoning classification are closely related while acknowledging that zoning regulations are only a proxy of actual land use imposing restrictions.

6.4 Common spatial representation

The first obstacle to studying the relationship between phone activity and land use is the reconciliation of the spatial dimensions of the data: While the location of the phone activities are recorded as coordinate pairs, zoning data is provided in polygons at roughly the parcel scale. The spatial partitioning of phone and population data is rarely the same as zoning parcels. To reconcile all data sources as well as to reduce the influence of noise (due to inter alia sources localization estimation noise) in the data, we transform both to the same uniform grid. A lattice is laid over the analysis region such that every cell in the lattice measures 200 by 200 meters. Different grid sizes have been tested. A size of 200 meters proved coarse enough to reduce the noise level and detailed enough in order not to mix many parcels of different zoning areas.

In order to reduce the high noise level average hourly time series of phone activity are computed. Here, the average is computed for each hour within a day of the week. Only cells with mobile phone activity above a certain threshold are used in the analysis. Similarly, the number of POIs tagged with each of the 104 unique descriptors is counted for each grid cell. With respect to zoning data, each cell is given a single zoning classification based on the most prevalent (in terms of fraction of area covered) use within the area.

Potential pitfalls of this method arise due to large heterogeneity in population density. Downtown areas are much more densely populated than the suburbs, a characteristic that is reflected in other spatial divisions like census tracts. This leads to sparse mobile phone activity in rural regions. However, the small grid size used in this analysis retains detailed information about block to block zoning regulations in dense urban areas. Figure 6-2 displays actual zoned parcels versus the gridded approximations.

Table 6.1 shows the frequency of each zoning class in the grid. The vast majority of land, nearly 75% of cells, are zoned as Residential. Other uses appear in roughly equal fractions.

Table 6.1: Tabulation of Boston zoning. The land use profile of the city is dominated by residential use accounting for nearly 75%. Other uses share roughly the same percentage of remaining land.

Zone Use	Category Index	Count	Percentage
Residential	1	23322	74.28
Commercial	2	1854	5.90
Industrial	3	2236	7.12
Parks	4	1941	6.18
Other	5	2045	6.51

6.5 Descriptive Statistics

We next examine the relationship between mobile phone activity and land use at the macro, city-wide scale. Figure 6-3 displays time series of mobile phone activity averaged over all cells of a given zoning classification. Examining absolute counts of mobile phone events reveals that the average activity differs greatly between zoning classifications. While residential areas only show a maximum activity of roughly 50 events per hour, commercial cells reach approximately 100 events on average.

The spatial distribution of activity is also heterogeneous. The downtown area of Boston shows orders of magnitude higher activity levels than typical residential zones. In order to allow for classification based on relative mobile phone activity, time series are normalized using a z-score. By definition, the normalized time series have zero mean and unit standard deviation. Mathematically, the normalized activity of cell (i,j) is given by:

$$a_{ij}^{norm}(t) = \frac{a_{ij}^{abs}(t) - \mu_{a_{ij}^{abs}}}{\sigma_{a_{ij}^{abs}}} \tag{6.1}$$

The second row of Figure 6-3 (a) shows the average (over cells of one zoning class) normalized activity. These profiles are remarkably similar for all zoning classes

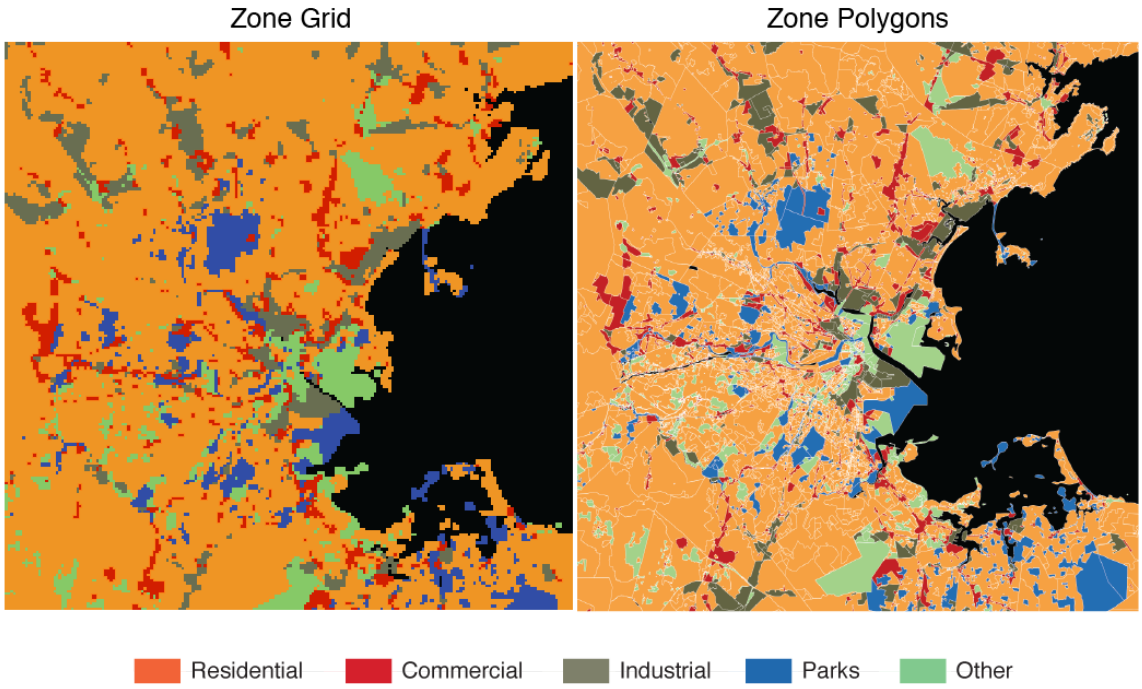

Figure 6-2: To improve computational efficiency and reconcile all mobile phone and traditional data sources, we create a uniform grid over the city. Zoning polygons (right), are rasterized to cells 200m by 200m in size (left). For cells where more than one zoning class exists, the most prevalent class is used. Given the small size of these cells, this data transformation provides an accurate map of the city while improving computational efficiency.

showing the strong circadian rhythm of the city. Residents wake up, go to sleep, and wake again the next day. The rise and fall of activity in each zone, however, is not solely the result of users moving into and out of a region. Instead, it is largely due to an uneven distribution of phone use across the day. To account for this, during each hour, we subtract the average normalized activity of the entire metro region from the normalized activity at each given cell. The corresponding spatially de-meaned series will be referred to as *residual activity*. Residual activity can be interpreted as the amount of mobile phone activity in a region, at a given time, relative to the expected mobile phone activity in the whole city at that hour. Mathematically, it is calculated as follows:

$$a_{ij}^{res}(t) = a_{ij}^{norm}(t) - \bar{a}^{norm}(t) \tag{6.2}$$

where $\bar{a}^{norm}(t)$ is the normalized activity averaged over all cells at each particular

time.

Averaging the residual activity for each zoning classification reveals patterns related to travel behavior. The last row of Figure 6-3 (a) and (b) provide the residual activity averages across zoning classes for weekdays and weekends. The most notable signal is the inverse relationship between residual activity in residential and commercial areas: While residential areas on average show higher than expected activity during the night and lower than expected during weekdays. As expected, the opposite is true for commercial zones. Somewhat surprisingly, the normalized activity does not show these features strongly. Only the residual activity demonstrates the expected behavior. There, also higher than average activity in parks on the weekend afternoons is visible.

Residential areas have higher residual activity in the early morning hours and late at night, while commercially zoned cells have a peak period during the day and show much lower activity levels late at night. These patterns most likely reflect the 9-to-5 business hours of offices and stores. More subtle patterns are also visible. In Boston, much of the CBD is zoned as Other or Mixed use. We see that residual phone activity in this zoning type has peaks in the early morning hours on Saturday and Sunday, suggesting these areas support night life on the weekends. These city-wide time series show that mobile phone activity and land use are linked at the highest level of aggregation. By treating phone activity as a proxy for the spatial distribution of people at a given time period the expected patterns of concentration of people in the CBD and inner city region during the working day, and the shifts induced by the commuting behavior are visible in the residual activity levels.

Figure 6-4 displays the spatial distribution of normalized activity (top row) and residual activity (bottom row) at three time instants. Not shown in the plots are the absolute activity levels which are distributed much like population density. The CBD of Boston has orders of magnitude more activity than the rest of the city. Mapping the logarithm of absolute activity over time once again only reveals the circadian rhythm of the city which strongly dominates the differences in land usage which consequently are not seen in these plots.

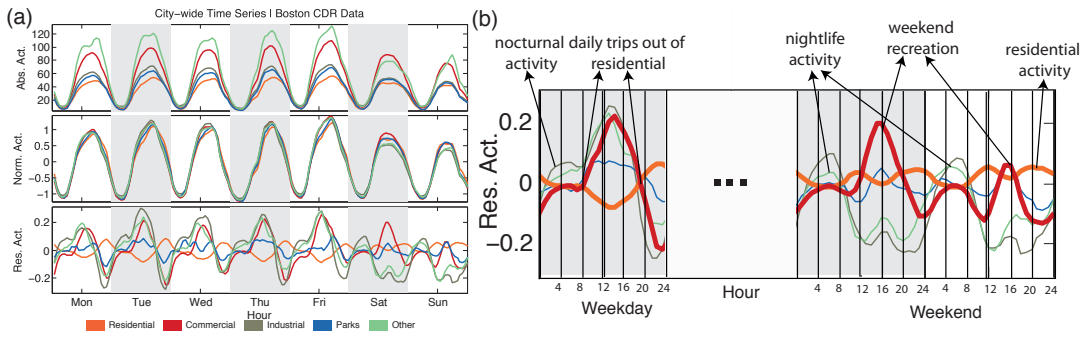

Figure 6-3: (a) Plots are shown for three different time series of average mobile phone activity within each of five land use. The first plot shows absolute activity (number of calls and SMS messages). The second plot displays z-scored time series. The bottom plot shows residual activity.(b) More detailed view of average (over cells of the same zoning class) residual activity.

In the spatial distribution of the normalized activity the dominance of the CBD is less pronounced. Nevertheless, the circadian rhythm still dominates the differences between different zones. From this perspective, Boston appears as a monocentric region, with small pockets of density located on an urban ring roughly 20km from the CBD.

By contrast, the spatial distribution of residual activity reveals a much richer structure. In the early morning hours, residual activity is located on the periphery of the region. During the day, this activity becomes heavily concentrated in the CBD or in small subcenters on the urban ring. Later in the evening, activity again returns to residential areas on the periphery, away from centers. This suggests some correlation between commuting patterns and the spatial distribution of residual activity.

In addition to the spatial distribution of mobile phone activity, we also explore point of interest density for the metro region. Figure 6-5 displays counts for the twenty most common POI tags as well as a spatial density plot showing their distribution in space. More general tags such as 'establishment' are featured prominently, while some tags like 'park' match very closely with official zoning classifications. As expected, we find that the CBD has the highest density of POIs, with smaller secondary centers visible as seen in mobile phone activity.

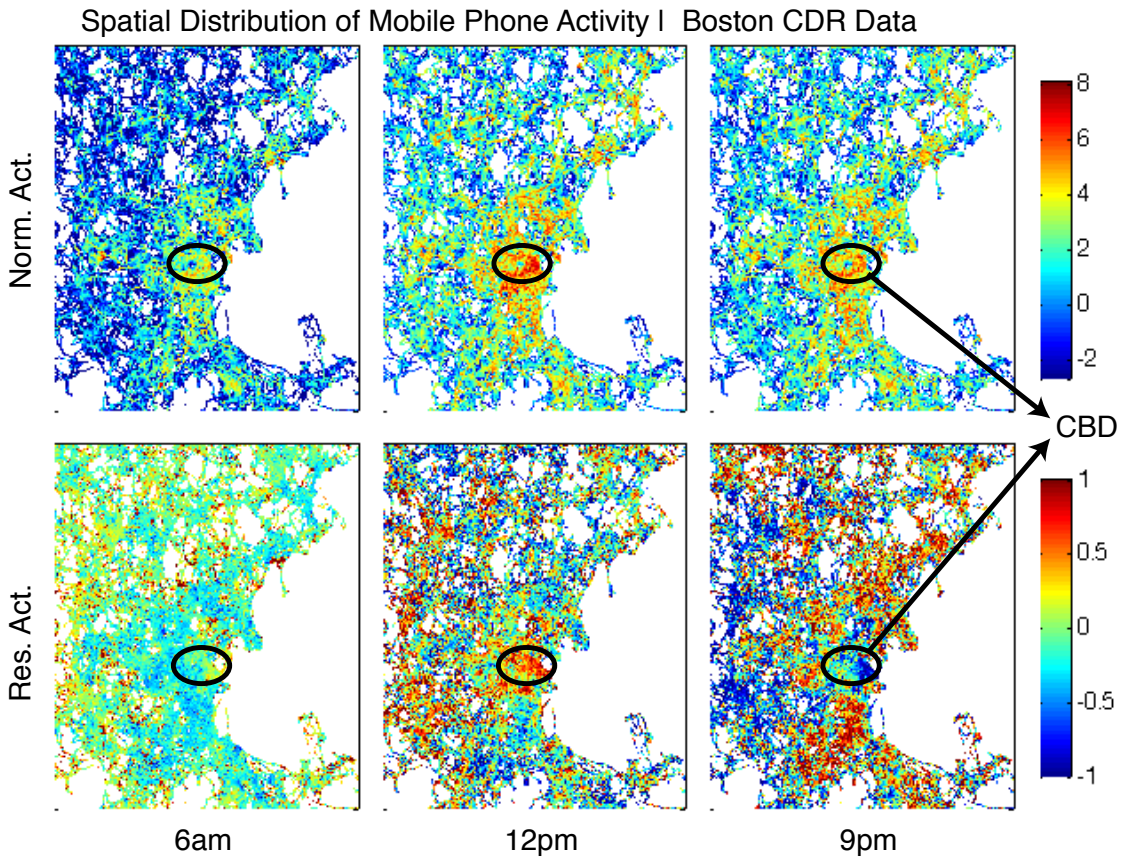

Figure 6-4: Spatial distribution of absolute and residual phone activity over the course of a day. While absolute mobile phone activity is dominated by population density and sleep and wake patterns, residual activity reveals flows into and out of the city center over the course of a day.

6.6 Classifying Land Use by Mobile Phone Activity

In the last section we observed correlation between residual mobile phone activity and land use on the macro scale. Fluctuations in mobile phone activity mimic our intuition of population changes related to commuting and recreational trips. In this section we investigate whether usage patterns in cell of a given class are homogeneous. This will be done by performing supervised classification based on features extracted from the residual activity time series and the classes provided by the zoning regulations as labels. Though previous work in this area has employed unsupervised learning techniques, access to extensive zoning data in a mature, regulated city such as Boston makes supervised learning an attractive option. Cross validation is used to

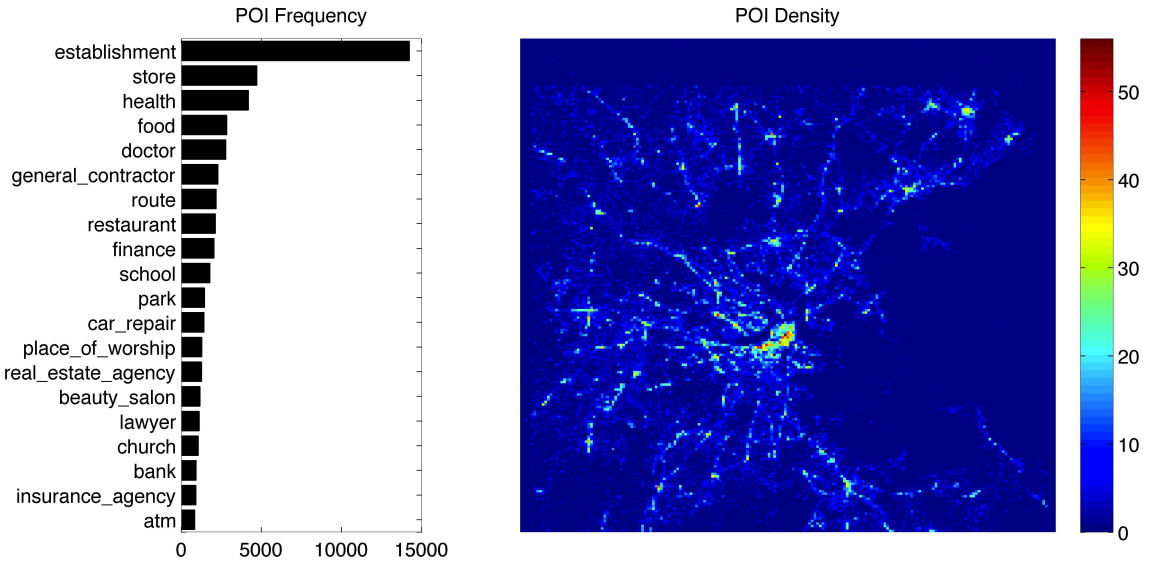

Figure 6-5: The left panel displays the region-wide frequency of the 20 most common point of interest tags. While some descriptors such as 'establishment' are vague, others like 'park' align closely with official zoning classifications. The right panel shows the spatial density of all POIs across the region. The central business district predictably shows the highest density, while other clusters are dispersed throughout the region.

test performance.

We implement the *random forest* approach described by Breiman [35]. Other approaches including neural network based classifiers have been tested and led to similar results. Random forests are useful for their ability to efficiently classify data with large numbers of input variables (such as long time series). Rather than make comparisons for every feature of the data every time, a number of random subsets are chosen to more efficiently search the space. This does not come at the cost of accuracy as random forests have been shown to have high performance on a variety of datasets [35]. Moreover, random forest classifiers allow weights to be introduced so that more frequently occurring classes do not overwhelm smaller ones. This feature will be exploited later to control for the large share of residentially zoned locations.

A random forest, $\{h(\mathbf{x};\theta_k), k = 1, ...\}$, is constructed from a set of decision trees as visualized in Fig. 6-6. The training data determines the parameter vectors θ^k. Least squares or maximum likelihood estimation can be used to find these configurations. To obtain a single prediction for each input time series, a voting scheme is implemented. Each tree votes for a class based on its prediction. These votes can be weighted

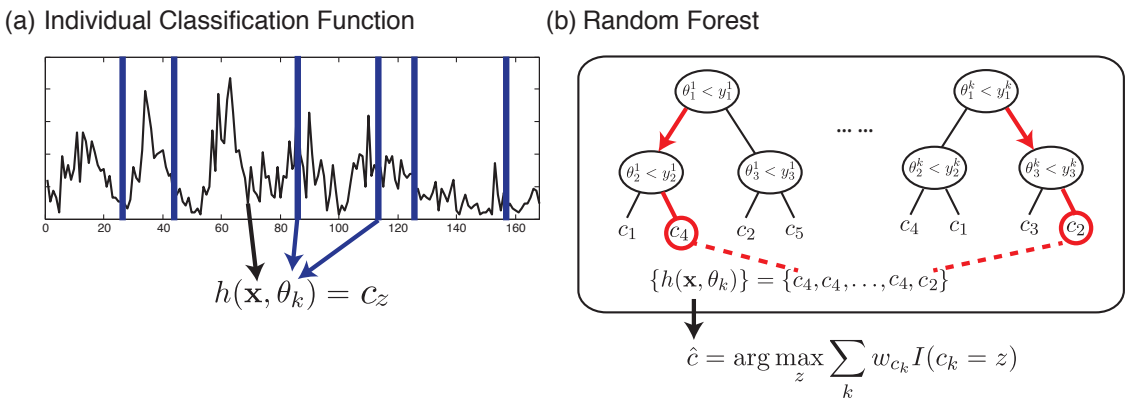

Figure 6-6: (a) Shows the inputs to each decision tree $h(\mathbf{x}, \theta_k)$. A time series of residual phone activity, $\mathbf{x}$, is input and activity at a random subset of times , θ_k (denoted by the blue bars), is chosen to make comparisons. (b) A depiction of the random forest shows a number of different trees making predictions based on a different set of random times. Each tree casts a weighted vote for a certain classification. A final classification, $\hat{c}$, is made by counting these votes.

(weights denoted by w_{c_k}) so that votes for one class count more or less than votes for a different class. The weighted votes are summed and a single zoning class prediction, $\hat{c}$ is chosen for the original input time series.

For the calculations we use a MATLAB implementation of the random forest algorithm released by Jaiantilal [2]. Given the periodicity observed in the data, our initial approach uses 49 input features which are computed for each location as the input feature vector $\mathbf{x}$. These features include a 24-hour time series of residual mobile phone activity during an average weekday as well as a 24-hour time series of residual activity for an average weekend-day. The final feature is the mean of the location's absolute activity on any given day. Additional features such as the variance of mobile phone activity were tested, but none aided prediction. The output of the algorithm is a zoning classification for each location. Cross validation is used to test accuracy. We create 500 trees for each forest and define total accuracy as the fraction of correctly classified cells on the validation part of the sample.

Our first set of results include all five zoning classifications: Residential, Commercial, Industrial, Parks, Other. When all land use classes are included, however,

[2] `http://code.google.com/p/randomforest-matlab/`

we face a major challenge with classification. As noted above, nearly 75% of all cells are primarily residential. The next most common zoned use is Industrial at 7%. Because of our definition of total accuracy, the most naive classifier, simply assigning Residential to everything, will achieve 75% total accuracy, but will fail to capture any diversity in use. To guard against this, we weight the voting system to raise or lower the required votes in order to choose a given classification. The maximum of the weighted votes then provides the predicted class. Systematic variations of the weights on a (coarse) grid led to a choice of weights where the criterion applied was maximum classification accuracy for all classes but residential.

Finally, we note that the random forest classifier uses local information only to make a prediction. Given the size of our grid cells, it is reasonable to assume that land use does not differ greatly from each 200m by 200m tract of land to the next. To incorporate neighborhood information into our predictions, we implement a second pass algorithm. After the classifier has made a prediction for a cell, we examine the predictions for each of that cell's neighbors. If the majority of neighboring cells were predicted to be a land use that differs from the cell in question, that cell is switched to the majority use of its neighbors. In practice, this results in some spatial smoothing of noisy classification data. We find that performing the second pass provides gains of 2-10% overall accuracy for each classifier.

Even with vote weighting and the second pass algorithm, we achieve only modest results. Table 6.2 shows 54% accuracy over the whole city. This implies that striving for equal classification accuracy among prevalent classes reduces overall accuracy by about 20%. Figure 6-7 displays the spatial distribution of correctly and incorrectly classified locations. We note, however, that the algorithm does capture some spatial patterns in the data and that our intra-use accuracy is relatively high for Commercial and Industrial uses. Parks and Other mixed uses remain difficult to classify.

To account for the tendency of the algorithm to over-predict residential use, we remove cells zoned as Residential from consideration. This leaves a nearly equal share of the remaining four uses: Commercial, Industrial, Parks, and Other. Table 6.3 and Figure 6-8 display results for this sub-classifier. Now, the zone with the largest share

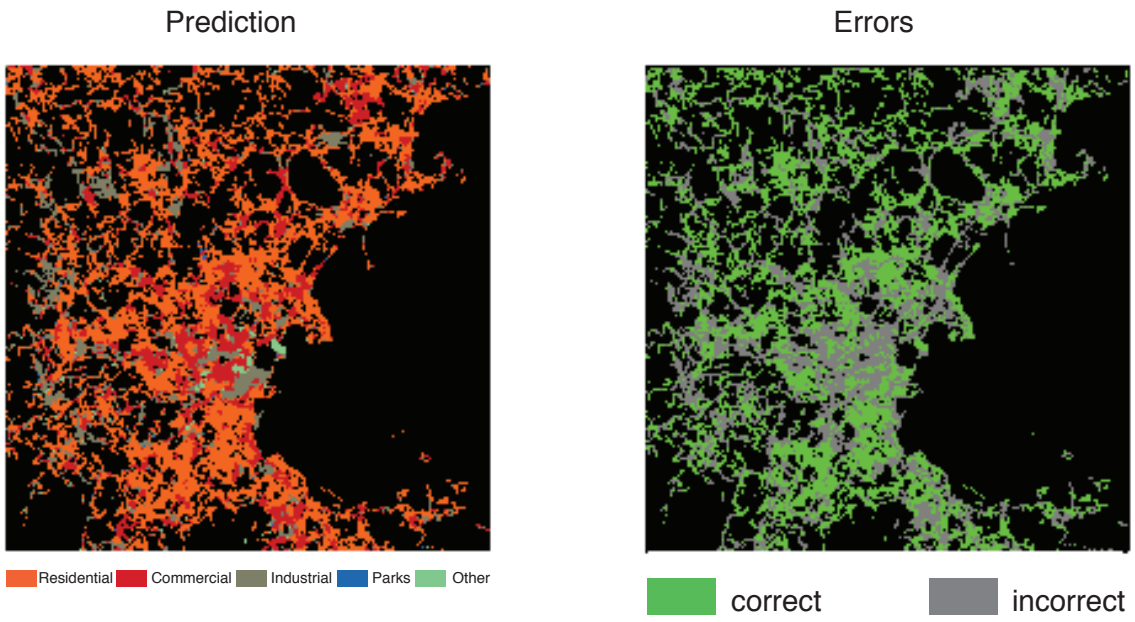

Figure 6-7: Left plot: zoning map as predicted from mobile phone data using the random forest classification algorithm. Right plot: spatial distribution of where the algorithm predicts land use correctly and where it fails. In general, these errors seem randomly distributed in space, suggesting that errors are not the result of some spatial correlations such as population density. For comparison to actual zoning, see the left panel of Figure 6-2.

is commercial use, which only accounts for 33% of non-residential zones. Intra-use accuracy has improved significantly for Parks and Other mixed uses. Whereas the random forest including residential uses could only correctly classify 2% of zones classified for Parks, the sub-classifier, excluding Residential, correctly predicts 30% of park cells. A similar improvement from 10% to 34% is also observed for the Other or mixed use category. The share of classes incorrectly classified as Residential roughly is distributed onto Parks and Others in the classifier without the Residential category, while commercial and industrial zones are not affected heavily. One hypothesis for this effect is that many cells while classified as Residential in rural areas are not fully developed and thus used as parks and in the city center show mixed usage. Including the large class of residential zones masks this effect.

The goal of the supervised learning algorithm is to make correct predictions of actual zoned use. Incorrectly classified cells are labeled as errors, but how an area is zoned is not necessarily the same as how it is used. As an example the area termed “Back Bay" containing some of Boston’s most busiest shopping streets is classified as

Table 6.2: Random forest classification results. The threshold refers the total number of phone events required in each cell over period of data collected to be considered for classification. Total accuracy is defined as the fraction of correctly classified cells. The share refers to the percentage of cells actually zoned for each class of use. Element (i, j) of the confusion can be interpreted as the fraction of actual zoned uses of class i that were classified as use j by the random forest. Thus the high percentages in the Res column can be interpreted as the algorithm heavily favoring classification as residential due to its overwhelming share of overall uses.

Total Accuracy:	0.54				
	Res	Com	Ind	Prk	Oth
Land Share:	0.74	0.09	0.08	0.04	0.05
Vote Thresh:	0.60	0.10	0.10	0.10	0.10
			Confusion Matrix		
	Res	Com	Ind	Prk	Oth
Res	0.62	0.21	0.15	0.01	0.01
Com	0.30	0.48	0.19	0.00	0.02
Ind	0.33	0.27	0.38	0.00	0.02
Prk	0.52	0.26	0.18	0.02	0.02
Oth	0.37	0.28	0.25	0.00	0.10

residential, as is the campus of MIT. Clearly these areas have a different usage than residential areas in the suburbs. A political and idiosyncratic process for setting and updated zoning regulations may lead to broad or unenforced development standards. In light of this, errors made by our classification algorithm may be due to incomplete zoning data rather than actual mistakes. To examine this possibility further, we analyze prediction errors more closely then introduce alternative data.

Figure 6-9 displays a detailed partitioning of classifier results. We compare average residual activity across three groups of cells: (I) All cells correctly predicted to be a given use. (II) All cells of another use incorrectly predicted to be the given use. (III) All cells of a given use incorrectly predicted to be some other use. Reviewing residential use, we see that Group I is defined as all residential cells correctly predicted to be residential. The average activity pattern is the most dominant pattern of residual activity for residential land use. We find that the residual activity in non-residential cells predicted to be residential (Group II) closely follows the pattern found in Group I. This strongly supports our hypothesis that though some zones are not classified as residential in the data, their phone activity patterns suggest they are

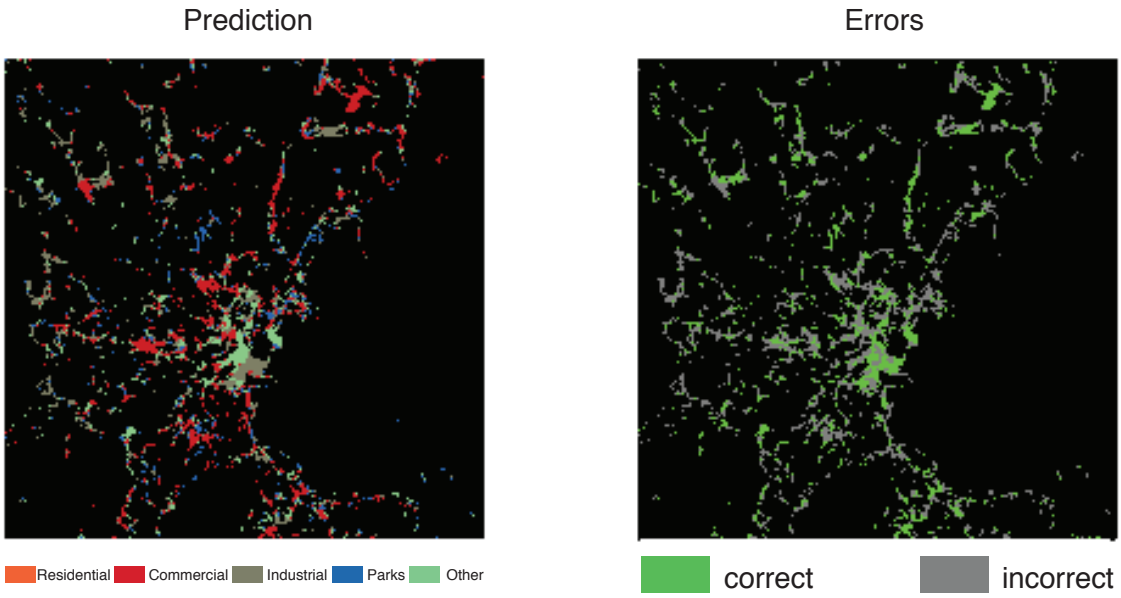

Figure 6-8: The left plot shows the city zoning map with residential areas removed as predicted from mobile phone data using the random forest classification algorithm. The right map displays the spatial distribution of where the algorithm predicts land use correctly and where it fails. Without residential areas to predict, the algorithm performs significantly better at predicting other uses. For comparison to actual zoning, see the left panel of Figure 6-2.

used in similar ways. In contrast, the residual activity in residential cells incorrectly classified as some other use (Group III) displays the inverse pattern. This suggests our algorithm is identifying cells that are zoned as residential use but that do not share activity characteristic of that zoning class in reality.

6.7 Incorporating Points of Interest

To further test our hypothesis that official zoning classifications may be incomplete, we incorporate point of interest (POI) data. Perhaps there is some behavior specific to when individuals use mobile phones that makes them unable to accurately proxy for activity in an area. POI data may provide a more direct measure of what is actually built in a region. With this in mind, we augment the feature vectors input into the random forest algorithm to contain both mobile phone activity and the composition of POIs in that grid cell.

Table 6.3: Random forest classification results - less residential. In this case, residential land has been removed from consideration. The algorithm is now able to correctly predict much larger fractions of rarer land uses.

Total Accuracy:	0.40				
	Res	Com	Ind	Prk	Oth
Land Share:	0.00	0.33	0.31	0.16	0.20
Vote Thresh:	N/A	0.30	0.30	0.20	0.20
			Confusion Matrix		
	Res	Com	Ind	Prk	Oth
Res	N/A	N/A	N/A	N/A	N/A
Com	N/A	0.50	0.19	0.11	0.19
Ind	N/A	0.27	0.37	0.12	0.24
Prk	N/A	0.31	0.18	0.29	0.21
Oth	N/A	0.26	0.24	0.15	0.34

For each cell, we create a feature vector with 104 elements (one for each unique tag) are counts of the number of POIs with that tag within the cell. These features may then be appended to the CDR data and input into the random forest algorithm.

To test the affect of adding POI information, we first attempt to predict official zoned uses using only the POI feature vectors, holding out mobile phone data. Total accuracy drops from 54% to 42%. Most of this decline is due to very high error rates in residential areas. Places which contain mostly private residences do not appear as places of interest on the map and are thus difficult for the algorithm to predict. Though overall accuracy falls, accuracy within other groups increases. Table 6.5 shows that success rates in commercial, park, and other/mixed areas rise by 8, 36, and 4 percentage points, respectively when compared to using CDR data (Table 6.2). These results are somewhat expected given the existence of tags directly indicating the existence of commercial establishments and parks.

With mobile phone data able to separate residential from non-residential zones and POI data better suited for classifying non-residential types only, we next combine the two data sources. The combination is performed in two ways. The first runs our random forest algorithm with augmented feature vectors containing both mobile phone activity and POI data as input. The second performs a two stage classification procedure. The two stage process first uses CDR data to predict residential or non-

Table 6.4: Random forest classification results - POI Only.

Total Accuracy:	0.42				
	Res	Com	Ind	Prk	Oth
Land Share:	0.74	0.09	0.08	0.04	0.05
Vote Thresh:	0.60	0.10	0.10	0.10	0.10
			Confusion Matrix		
	Res	Com	Ind	Prk	Oth
Res	0.42	0.16	0.18	0.17	0.08
Com	0.09	0.54	0.27	0.05	0.06
Ind	0.15	0.25	0.42	0.09	0.10
Prk	0.24	0.14	0.12	0.38	0.12
Oth	0.19	0.29	0.16	0.22	0.14

residential then takes all cells predicted as non-residential and classified them again as Commercial, Industrial, etc. using only POI feature data. Both methods give similar results. The addition of POI information increases accuracy by roughly 15 percentage points to 70% compared to CDR data alone. Moreover, improvements are made in mostly in commercial regions as apposed to residential.

Table 6.5: Random forest classification results - POI and CDR - One and Two Stage

One Stage						Two Stage					
Total Accuracy:	0.68					Total Accuracy:	0.71				
	Res	Com	Ind	Prk	Oth		Res	Com	Ind	Prk	Oth
Land Share:	0.74	0.09	0.08	0.04	0.05	Land Share:	0.74	0.09	0.08	0.04	0.05
Vote Thresh:	0.60	0.10	0.10	0.10	0.10	Vote Thresh:	0.60	0.10	0.10	0.10	0.10
	Confusion Matrix						Confusion Matrix				
	Res	Com	Ind	Prk	Oth		Res	Com	Ind	Prk	Oth
Res	0.79	0.15	0.05	0.01	0.01	Res	0.89	0.05	0.01	0.02	0.03
Com	0.27	0.61	0.11	0.00	0.01	Com	0.58	0.31	0.05	0.02	0.05
Ind	0.35	0.30	0.34	0.01	0.00	Ind	0.54	0.23	0.13	0.03	0.07
Prk	0.66	0.22	0.06	0.03	0.03	Prk	0.72	0.09	0.02	0.09	0.07
Oth	0.51	0.27	0.13	0.01	0.08	Oth	0.56	0.23	0.05	0.05	0.11

We note that significant portions of the city remain difficult to classify. The limited impact of additional detailed and current point of interest data provides further evidence that official zoning regulations does not guarantee uniform patterns of activity.

6.8 Conclusion

In this chapter, we examined the potential of CDR data to predict land usage. We demonstrated that aggregate data shows the potential to differentiate land usage based on temporal distribution of activities. While the absolute activity is dominated

by the circadian rhythm of life, eliminating this rhythm reveals subtle differences between the five main land use categories Residential, Commercial, Industrial, Parks and Other. The addition of a temporal dimension to zoning classification may aid strategic planning decisions related to land use.

As the data are available at a high spatial resolution, we investigated the capabilities to infer land use on a fine grid of 200 by 200 meters. We found that supervised classification based on labeled zoning data provides estimated land use classifications which show better accuracy than random assignment. At the same time accuracy is worse than classifying every zone as Residential, the dominant category.

Reasons for this lack of accuracy might be found in the nature of the data used: actual usage might differ from the zoning regulations and Residential is often confused with Parks and Other zones. Omitting residential zones, the classification accuracy for Parks and Other zones greatly increases while industrial and commercial zones classification accuracies are not heavily affected. For rural areas where residential land might not be fully developed this is plausible. For urban zones the distinction between Residential and Other zones might also be subject to temporal changes as mixed use is prevalent. Finally, analysis of prediction errors reveals that the algorithm fails to correctly classify areas because they have fundamentally different mobile phone activity patterns. This suggests that there may be heterogeneity in how land is actually used, despited its official zoned classification.

To investigate this further, we crawled and incorporated additional point of interest data for the region. These points of interest present a much more detailed and current picture of what type of businesses and services exist throughout the city. However, we find little correlation between POI data and official zoning classifications. Though predictions of zoning types other than residential are improved modestly using POI data, the mapping is far from perfect. A hybrid approach, using both mobile phone and POI data as feature vectors in classification provides the best performance. In total, we are able to correctly classify roughly 70% of locations in a city, provided enough call data is available. Classifying zones other than residential and commercial remains challenging. These results suggest that official zoning

classifications fail to sufficiently describe how an area will be used. Further work will explore the relationship between mobile phone activity and points of interest more closely, leaving official regulations behind.

Thus the main conclusion is that the CDR data shows some potential to infer actual land use both on an aggregate level and on a higher spatial resolution. However, zoning data might not be the optimal data source to infer actual land use and hence act as ground truth to guide the supervised learning algorithm. In this respect, our analysis suggests that mobile phone activity may be used to measure the heterogeneity in how space is used that cannot be captured by simple and broad zoning classifications. Moreover, the incorrect predictions made by our algorithm with the addition of alternative data may suggest updates to traditional zoning maps so as to better reflect actual activity or highlight areas where more planning oversight is needed.

Collectively, these results provide a tool that can be used to augment static measures of population distributions with high resolution spatiotemporal dynamics. We hope this information will be useful to make effective and efficient choices of locations for both public and private resources. In addition to potential applications, we hope that tools and techniques developed and applied above will prove useful to merging traditional and novel data.

6.9 Acknowledgements

The work in this chapter was the result of a collaboration with Michael Ulm, Dietmar Bauer, and Marta C. González. This work was made possible in part by the MIT - Xerox Fellowship as well as an National Science Foundation Graduate Research Fellowship.

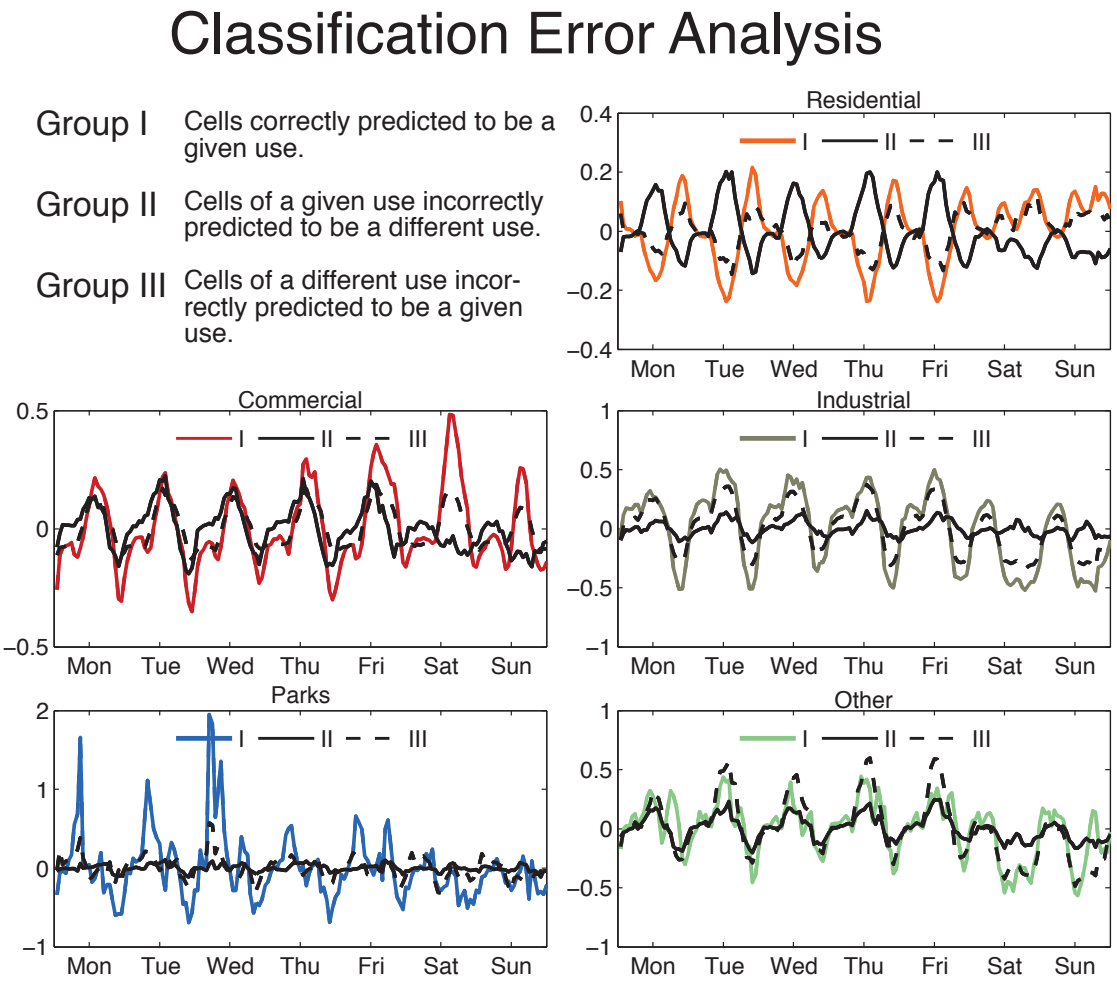

Figure 6-9: An analysis of classification errors. We consider three groups: (I) Cells correctly predicted to be a given use (II) Cells of a given use incorrectly predicted to be some other use (III) Cells of some other use incorrectly predicted to be a given use. For example, Group I includes all residential areas correctly predicted to be residential. Group II, residential cells predicted to be some other use (i.e. Commercial), have average activity that is the inverse of Group I, suggesting these locations were misclassified because they display fundamentally different activity patterns. Group III represent cells of other uses such as Commercial that behave like Residential.

Chapter 7

Failures in markets for personal data assets.

7.1 Introduction

In March 2014, a study was published in the Proceedings of the National Academy of Science (PNAS) titled "Experimental evidence of massive-scale emotional contagion through social networks."[127]. The study, a collaboration between researchers at Cornell University and Facebook's internal data science team, used Facebook's enormous online social network to find empirical evidence that "[e]motional states can be transferred to others". To test this hypothesis, Facebook conducted a randomized, controlled trial to assess whether or not a Facebook user would change what they posted on the website in response to experiencing posts of various emotions from their friends. The results confirmed the hypothesis that emotions can spread just like disease or gossip and the extremely large sample size ($N = 689,003$) demonstrated the enormous potential to use online platforms to study social phenomena. Individuals who were shown negative content created by their friends were more likely to post negative content in return, an undoubtedly interesting finding. The problem, however, was that participants in this experiment were never informed that they were human research subjects. The media immediately seized the story, running sensational headlines such as the Huffington Post's "You May Have Been A

Lab Rat In A Huge Facebook Experiment"[100] and Slate's "Facebook's Unethical Experiment"[226]. Academics, competitors, and ethicists followed soon after with their own take on the matter.

While Facebook's experiment was perfectly legal, scientifically valid, and approved by the Independent Review Board (IRB) charged with overseeing academic research on human subjects, it was a spark that ignited an intense debate over the use of online platforms and the data they collect on users for research. It exposed deficiencies in the legal, ethical, and scientific frameworks to deal with rapidly evolving data and methods. As private entities like Facebook are scrambling to expand, organize, secure, and ultimately leverage data to reduce costs, improve efficiency, and add value to products and services, we must address these issues. The World Economic Forum suggests data is the "new oil of the 21st century" and its economic potential is estimated to be in the hundreds of billions of dollars[145]. The categorization of data as an economic asset is more than a simple cliche, it provides a framework to study issues related to privacy, safety, and property rights.

Though the economic benefits to companies are potentially vast, the dramatic increase in size of personal data presents very real threats to privacy, safety, and property rights of consumers. Data security breaches can release hundreds of millions of names, addresses, and credit card numbers in an instant, making identify theft possible on a massive scale[30]. Compounding the damage, massive negative externalities are generated when breached data is used to infer things about the friends of individuals who may not even know they are at risk. As online and mobile technologies play a more central role in cultural movements like the Arab Spring, governments have large incentives to track and control data on citizens. At the same time, the ability to collect new data and perform massive experiments on users to improve or change products raises numerous questions of the ethical practice of research. These experiments are no longer confined to academic or public institutions with well established protocols and safeguards and the definition of what is considered human subjects research is blurred by the speed at which experiments are conducted by and turned into consumer products.

As the value of personal data increases, questions of who has rights to collect, use, and provide or receive compensation for this data are only becoming more important. Companies invest in and maintain capital, products, and services that are offered for free to users in exchange for the right to sell data gathered to third parties such as advertisers. Consumers are becoming increasingly aware of and concerned about these practices and are seeking protection.

The Government's role in big data is equally complicated. Public institutions from the municipal to federal level have long been collectors and curators of massive data assets. In the United States, the federal government is required by the Constitution to conduct a census that includes every citizen. State and local governments collect data on variables from health insurance claims to criminal offense reports. While these data collection efforts are relatively transparent and understood, they pail in comparison to the alleged covert data collecting activities of intelligence agencies such as the NSA and CIA. These institutions are under increasing pressure to make data available and practices transparent to citizens. In addition to collecting data themselves, governments are also charged with protecting citizens. With the concerns over safety and security surrounding personal data, more people are calling on the government to regulate the collection and use of data by private companies. Governments now face the challenge of finding a solution that respects the interests of citizens without stifling innovation and progress in the private sector.

This chapter is divided into three sections. The first describes three sources of failure in markets for data assets, uncertain property rights, information deficits and asymmetries, and externalities. The second outlines a number of private technical and self-regulatory solutions to these failures, though none of these options emerge as a cure all. The final section discusses regulatory frameworks in existence or being developed by the United States and European Union. Though the guiding principals of their strategies differ, they differ on implementation.

7.2 Market Failures

Data is being collected, stored, and used at massive scales. The immediate economic benefits from leveraging data assets have created a figurative gold rush to uncover new efficiencies, personalized products, and new revenue streams with information. However, the diffusion technologies to capture and techniques to process data has out paced mechanisms to correct for a number of market failures including uncertain property rights, information asymmetries, and externalities. These failures threaten the long-term health of data markets.

7.2.1 Uncertain Property Rights

Property rights are central to efficient markets. More specifically, in the case of intellectual property such as data, they incentivize investment by ensuring costs of collecting and analyzing information can be recovered later. One way to ensure recovery is by creating artificial scarcity in the way of government granted monopoly rights, thus driving up prices and profits of rights holders. On the other hand, social welfare may be maximized by the diffusion of innovations so that they may be built upon further. Data, however, presents unique challenges when balancing these two competing objectives.

The World Economic Forum's 2012 report on "Rethinking Personal Data" identifies three classes of personal data: volunteered data, observed data, and inferred data. Volunteered data is information individuals choose to disclose, such as a blog post or photo. Observed data is data generated "as the result of a transaction with a company." For example, the time and location of a mobile phone is recorded when a text message is sent because the user is billed for this action later. Finally, inferred data is new information derived from some original data. Raw data can be aggregated to produce insights that were not apparent from individual record.

Data is a non-rival good. While it can be easily copied, transmitted, and used by multiple parties without preventing anyone else from using it, collecting data in the first place sometimes requires substantial investment. This creates an incentive to

free ride off the data collected by others. Free-riding incentives are stronger for some types of data than others. Volunteered data is a raw resource, needing only a system to capture and store it. Observational data requires a system to actually do the observing. Finally, inferred data is derivative, created as the result of many hours of research or substantial computational resources. Rights claims and incentives related to raw data differ greatly from those of derived data.

Personal data also exhibits network externalities. Data related to whom an individual calls with their mobile phone may offer some insights into an individual's social behavior. When the time and location of those calls is known, however, it becomes possible to identify the nature of those social relationships (i.e. social versus work related). The value of aggregating multiple pieces of data is often greater than the value of each piece individually. Similar effects exist with respect to scale. Disaggregate data on an individual customer's movement in a city might inform targeting advertising, but aggregated data on millions of individuals may help redesign an entire region's transportation system, fight diseases like Malaria, and unlock new insights into human behavior. Though individuals may claim ownership of their single piece of the aggregate, who claims ownership on the value created by aggregation?

A company like Facebook spends large amounts of money to build a product to perform this aggregation only to offer the service to users for free. Of course, free is a misleading term. By design, this product collects data volunteered by users. These data assets are often used to improve the product itself, such as providing more relevant updates on your friends using the service. Ultimately however, a company may decide that it is necessary to monetize its data assets to recover fixed and operating costs. They may begin making inferences from volunteered data about demographics or buying habits that are then sell to third parties. Trouble arises when monetization is done in ways that users find inequitable or uncomfortable. Users of these products see this data as an extension of their own intellectual property, while companies see the inferences they created as new property they are entitled to.

Competing claims can be framed as differences in the interpretation of derivate works. With other intellectual property, such as copyrighted material, rights holders

are entitled to derivative works, or works based upon pre-existing intellectual property. If inferred data is a derivative of volunteered data, users may claim ownership. Companies see inferred data as transformative, giving them ownership of this new material. Though a useful categorization, words such as “derivative" and “transformative" works are not the words usually chosen. The World Economic Forum notes that rights involving personal data are operationalized by permissions. Permissions are licenses granting access to a piece of property for a specific use. Permissions can be temporary and depend on context. When a user signs up for a web application, they do not give up the rights to the data they volunteer, but rather sign an agreement allowing a private entity permission to use that data for certain purposes. The central debate is whether the creation of inferred data is allowed and, if so, who has rights to it. In theory, both parties have legitimate claims. Volunteered data does not exist without a company to build infrastructure or users to create content. In practice, however, consumers have overwhelmingly chosen to accept terms that give extremely broad rights to companies in exchange for a free service.

The current model is to present potential users with a permissions ultimatum at the time they sign up for a product or service. If an individual wishes to use a service, they must grant permission for a company to collect and use personal data. This often requires the user to read complicated terms of service (TOS) agreements and privacy policies written in opaque legal language. Though legal precedent has not yet been established in the context of personal data, it is not even clear that consumers can provide informed consent given the complexity of current TOS. When a company overreaches regarding the permissions they require or appear to have violated their own TOS, consumers generally have two forms of recourse, tort action and public pressure.

For example, in early 2009, Facebook quietly changed it’s terms of service agreement to include the clause that users “hereby grant Facebook an irrevocable, perpetual, non-exclusive, transferable, fully paid, worldwide license ... any User Content[.]"[196] Protests, some ironically shared on Facebook itself, spread rapidly as news organizations quickly picked up this story. This change sparked a debate over data sharing

terms of service on web applications. In a response posted on the Facebook blog, founder and CEO Mark Zuckerberg explained that their "philosophy is that people own their information and control who they share it with. When a person shares information on Facebook, they first need to grant Facebook a license to use that information so that we can show it to the other people they've asked us to share it with."[50] The language of Facebook's terms of service was eventually softened. As of February 2015, Facebook's TOS state that "you grant [Facebook] a non-exclusive, transferable, sub-licensable, royalty-free, worldwide license to use any IP content that you post on or in connection with Facebook (IP License). This IP License ends when you delete your IP content or your account unless your content has been shared with others, and they have not deleted it."[4] In addition to making permissions temporary, Facebook has also implemented more extensive privacy controls to limit what data they can provide to third parties.

In light of government failure to adequately define what companies can and cannot do with data, it is becoming increasingly popular to use tort action as a way of setting precedent. For example, in early 2012, a class action lawsuit was filed against 18 companies alleging theft of personal data from users' mobile devices[158]. According to the plaintiffs, mobile applications were illegally taking identifying contact information from address books then selling this information[142]. This argument implicitly follows from the belief that such data belongs to the user and that the user is entitled to value derived from it. Complicating matters further is the inclusion of handset manufactures such as Apple as defendants. Though Apple itself is not accused of stealing information, they are accused of facilitating theft by providing access as a middleman between application developers and users. If the plaintiffs in this case win the considerable damages they claim companies will undoubtedly re-think their roles in collecting data or even facilitating its creation.

In defense of claims over data, companies will note that there is nothing to force an individual to agree to these terms or use their products. However, as these services become more a part of cultural norms, consent seems more compulsory. At the same time, barriers to entry continue to rise as network effects make switching from one

service to another difficult. It is now necessary to consider the fairness of placing burdens on individuals for their choice not to be tracked. Furthermore, in business models where users may pay to retain data permissions, equity issues arise with those who cannot afford their own privacy.

Property rights and permissions also define what relationships companies can have with third parties. A major concern of consumers is that their data will be used for purposes entirely unrelated to the original context within which they provided it. For example, GPS manufacturer TomTom recently announced a partnership with UK auto-insurance provider Motaquote to offer discounted insurance to customers that allow TomTom to track their driving behavior[103]. While opt-in programs such as this provide reasonable levels of control, data sharing with third parties is often far less transparent. Consumers are rarely given the choice to approve of sharing in specific contexts, but are required to provide unrestricted permissions such as in the case of Facebook.

A final practical point surrounding data rights and permissions is the so-called "right to be forgotten." When a user severs his relationship with a company, what happens to his data? Can the company continue to use that data? Can the data be transferred to another individual? This issue is most salient in the case of death. With so many assets being digitized and managed online (e.g. books or photographs), families of the deceased often fight long battles to obtain usernames and passwords to access this data. Individuals are now advised to designate legal access to this data in their wills. Several commercial solutions have emerged promising to pass access to a designated recipient in the event of death. The ability to bequeath valuable assets after death is not new to intellectual property. Copyrights can last up to 90 years following the death of the creator. The length of data retention and permissions may have significant effects on incentives to provide or collect it.

In summary, property rights involving data are uncertain and complex. Data is often costly to collect. It is also a non-rival good that can be copied and transmitted at low cost. This creates incentives to free ride on those who invest infrastructure to collect it. As such, companies argue that in order to make investments viable, they

must be guaranteed monopoly profits. However, data is often derivative, creating conflicting claims in cases where substantial work has gone into transforming the raw data belonging to another party. Without balancing these concerns, companies may underinvest while consumers under-share. Uncertainty also exists for middlemen that control platforms. If Apple can be held liable for the actions of data collectors who use the iPhone as a platform for applications, they may significantly change which services are offered. Finally, the practical limits of data rights in cases such as death are fuzzy, often imposing emotion or monetary cost to settle. These battles are currently being fought via public relations or tort action.

7.2.2 Information deficiencies and Asymmetries

Even when property rights are well defined, both parties must be informed of those rights in order to make economically efficient decisions. This requires consumers to invest time and sometimes money to educate themselves before making choices about what to share and with whom. For example, consumers may be reluctant to use beneficial services for fear that their data will be used in some nefarious way. If a user attempts to educate himself on a company's data policies, he is faced with the challenge of navigating dense legal language. The length and number of terms required for all the services used on a daily basis summed over all individuals in an economy results in a huge time investment. A recent estimate places the opportunity cost to the United States economy of reading privacy policies as high as $781 billion annually[148]. As a result of these market conditions, consumers may choose to remain uniformed, leading them to make inefficient choices with regards to their data assets.

Consumers may not understand or foresee consequences of providing data to third parties. Moreover, they have no way of learning which third parties obtain their data. If they had this information, they may decline to provide data to the first party from the start. Similar over sharing may occur if users are informed of TOS, but not aware of substantial negative externalities that may arise from data sharing.

With most attention given to what data first and third parties can access, pro-

viding data to the user is often overlooked. An individual may volunteer valuable personal data to a company, but have no way to obtain or use the data himself. Beyond the utility a user might gain from using that data, access is vital to ensuring accuracy. As companies find new ways to utilize data to offer personalized products and services, ensuring the accuracy of information is a high priority. For example, though an advertiser uses browsing history to infer socioeconomic or demographic characteristics for use in targeted advertisements, individuals have no mechanism to even learn what the advertiser has inferred, much less correct an incorrect inference. Even if data is accurate, the opaqueness of TOS agreements often leaves users unsure of what third parties are accessing their data and why. This creates information deficits and that can lead to suboptimal market outcomes.

Consider an online retailer like Amazon. As a user browses, purchases, and reviews products, Amazon collects data on which items a users browses and purchases so that it can make inferences with respect to that person's preferences and characteristics (i.e. income). Ultimately, Amazon wishes to use these inferences to sell more things to its customers. If Amazon makes an incorrect inference, they may offer useless suggestions of products that individual would want to by. As a result, users remain unaware of potentially beneficial products and Amazon forgoes profit.

In the above scenario, it seems harmless if Amazon's use of a data recommends useless products to customers. Similar approaches such as A/B testing are standard practice. Companies will make multiple versions (version A and version B) of the same webpage or interface and monitor which versions generate the most clicks or purchases. Google, for example, is famous to testing over 40 shades of blue to see which resulted in more user clicks[33]. Visitors to these websites are not informed that they are part of an experiment and, in general, it would be difficult to argue that companies must only improve their product through guessing because collecting data is unethical.

As companies become more sophisticated in their attempts to improve services for customers, however, the harmlessness of these experiments. In the case of the Facebook experiment on emotional contagion, the results of the experiment showed

that a user of Facebook who was shown more negative posts will then post slightly more negative content themselves. The problem, now, is that in order to discover this, Facebook had to actually make some of its users sadder and they did so without the users knowledge. From a scientific perspective, it is misleading at best to suggest Facebook intentionally manipulated the emotions of users. The point of running an experiment is that the outcome is unknown before you actually run it. It is well known and advertised that Facebook curates the content shown to users and for all anyone knew, the content they were already showing users was inducing negative emotions. Moreover, Facebook did not manipulate the content of posts, it merely displayed those with slightly more or less negative words.

That said, it is easy to imagine the case where such an experiment could have disastrous consequences for unsuspecting users. For example, an individual suffering from depression may have opted out of participating in an experiment where they would risk worsening their condition. Anticipating these potentially negative impacts on human subjects, academy has developed a full array of guidelines for performing such research and were a similar experiment have been run at a university, fully informed consent of each participant would be required. Facebook, however, as a private entity whose users agree to terms of service allowing it, is not burdened by these structures.

As Michelle Meyer, former editor of the Harvard Law Review and professor of Bioethics at Union Graduate College, pointed out in her analysis of the scandal[153], the study most likely constitutes human subjects research and would require IRB approval were it subject to federal research regulations (it is not). Still many denounced the researchers at Cornell for help designing and analyzing the study. The problem, now, is the unintended consequences of this backlash. It is perfectly legal (and will likely remain that way) for private companies to perform these experiments and in many cases does create better products. Facebook's biggest mistake, it seems, was to make the results of these experiments public to the scientific community and the general population. With so much to gain from the results, Facebook's (or any company's) strategy should be to perform these experiments, but keep them secret.

To stop the flow of knowledge being cut off, some academics now residing in private research labs have offered a middle ground, whereby private corporations voluntarily submit proposed experiments to a review board with the promise that the process is more streamlined and less bureaucratic than what might be found in a university[233].

These recent incidents serve to highly the vast information asymmetries that exist where these new data sources reside. The generators of the data rarely have the time or the ability to understand he terms put forth by the collectors and perceived misuses threaten to keep research locked away as trade secrets. Both of these asymmetries only further the potential for market failures and inequitable outcomes and more must be done to reduce this risk.

7.2.3 Externalities

Even when markets have well defined property rights and full information, externalities can still lead to inefficient outcomes. Costs not imposed directly on producers lead to an over supply of some output, while benefits not directly capture lead to a scarcity. Data assets give rise to a number of these externalities both positive and negative. While companies gain value from using and selling data, they do not incur direct costs when that data is leaked or used in ways that harm the individuals it came from. At the same time, there are potentially massive public benefits from leveraging data assets collected by private companies. To complicate matters further, these externalities can be conferred onto individuals who choose to abstain from data sharing completely.

Google recently found itself facing backlash due to negative externalities spawned by their failed social network application, Google Buzz. In early 2010 Google used address book data from subscribers to its email application (Gmail) as a way “seeding" the Google Buzz network. Gmail users who enabled Buzz were automatically connected with other Gmail users they had contacted. These connections were then made public to other Buzz users. Adopters were not given the option of consenting to or filtering which connections were made or what data was shared publicly in their Google Buzz networks. This “feature" generated a number of negative consequences

for users such as revealing affairs or leaving individuals in tumultuous geopolitical regions exposed to tracking by government (or extra-government) agents[112]. Google acted quickly to correct these issues, but for many the damage was done. Finally, in late 2010, Google settled a class action lawsuit brought by disgruntled users, agreeing to pay $8.5 million to educate individuals about privacy on the web[70]. Google was forced to internalize some costs of privacy through loss of market share, as angry users chose alternative services and eventually paid a small sum settlement.

In 2011, the Internet activist group Anonymous claimed responsibility for hacking into Sony's PlayStation Network (PSN) and compromising more than 75 million usernames, passwords, and in some cases encrypted credit card information. The damage of data breaches in these cases is magnified by the tendency of individuals to use similar usernames and passwords across multiple accounts. An attacker needs only to identify the weakest point across all applications to gain access to everything else. While Sony has stated the direct cost of their data breach as topping $171 million, they have yet to compensate individuals or companies whose data may now be at risk due to reuse of credit cards, usernames, and passwords[195]. Furthermore, the most direct vehicle used to impose these costs on companies is tort action. This may lead to reactionary decisions by firms where data isn't secured until after an attack occurs. A more efficient policy might impose immediate, direct costs rather than rely on firms hedging against future legal threats. In practice, users also bear some responsibility for choosing strong and unique passwords and usernames to limit the damage. This, however, increases the burden placed on consumers as the number of applications requiring identification grows.

In addition to security concerns over data breaches, many online applications are used heavily by children. In these cases, there is an immediate question as to whether someone who is underage (13 years old in the US) can legally give permission to personal data access. Moreover, when a minor reaches a legal age of consent, it remains unclear what should happen to data generated by their younger selves. In the judicial system of the United States, for example, court records of juveniles are sealed even for those convicted of crimes. These laws are meant to shield children from

discrimination later in life due to acts perpetrated before they were legally capable of bearing responsibility. As employers become increasingly insistent on vetting the online presence of potential hires, these issues will become more prominent.

Governments do not always see an interest in protecting data privacy. In some cases, they actively seek to erode it. China's "Great Firewall" provides strict control over what online content citizens can access as well as and close surveillance of users. Google has recently spared with China over these policies, suspending service due to concerns over censorship. Fights like the one between Research in Motion (RIM), maker of the popular Blackberry phones, and Saudi Arabia over providing the government with access to encrypted messages will only become more important in the wake of Arab Spring movements, facilitated by mobile communication[156]. Even in places like the United States, law enforcement is increasingly using data from mobile phones for surveillance and evidence without warrants[177].

Concerns over the governments use of this data exploded after security consultant Edward Snowden leaked an enormous number of classified documents detailing shockingly extensive data collection and mining projects within the NSA and CIA of the United States[147]. While it may be reasonable to assume that governments maintain covert data collection operations on foreign entities, the true shock was the scale and scope of their data collection efforts on US citizens. Charged with indexing every email, text message, and phone call made by every person in the United States, these revelations spurred a furious debate. The government contends that it must collect this data to protect its citizens while many citizens feel their basic rights to privacy have been violated. Irregardless of ones position on these core issues, it is undeniable that these revelations have imparted large externalities on many parties. Large US tech companies have spent billions encrypting their systems and responding to charges of being complicit in handing over data to the government and have lost billions of dollars in contracts due to concerns that the US government will have access to sensitive information[155]. Moreover, there have been reports that, despite safe guards, these intelligence capabilities have already been abused to share personal information about innocent individuals[189]. Even foreign relations have been

damaged due to revelations that phone calls between diplomats were being covertly recorded[165]. The full implications of this information is still unknown, but the externalities from it is already being felt.

Not all externalities resulting from data use are bad. Data also presents an enormous opportunity to solve public problems on a massive scale. Locations recorded from millions of mobile phone users in a city may be used to relieve congestion, make public transportation more efficient, or uncover life saving responses to disease outbreaks. Digital health records could dramatically speed up drug discovery and reduce the cost of clinical trials[241]. However, in order for these benefits to be realized, companies need incentives and permission to collect and share data as well as an incentive to do so. Individuals must be informed of both the risks and rewards that arise from sharing their data.

Finally, externalities associated with personal data assets may be imparted on individuals who choose not to share. For example, recent research has demonstrated that it is possible to predict the locations visited by an individual knowing only the movement patterns of that individual's friends[78]. Even if a person does not give an enterprise permission to collect or observe data, they may still be at risk via friends who have done so. Extra effort should be taken to protect individuals who may not even be aware they are participants in data markets.

These scenarios highlight externalities inherent the collection and use of personal data. When companies do not incur costs associated with data and privacy breaches at the time they build their systems, they may take inadequate security measures. At the same time, there is little incentive to collect and use data to provide public benefits when those benefits do not appear on the balance sheets of a business. Currently, public relations and tort action have been the most widely used vehicles to impose costs on companies in the case of negative externalities. While governments have, in some cases, stepped in to mitigate negative externalities, some have also shown a willingness to forego privacy and security of citizens for control and political stability.

The three market failures described thus far present real threats to the formation and efficiency data markets. Moving forward, a combination of private and public

solutions may be necessary to ensure that markets continue to exist and function in more optimal ways. The remaining sections of this paper discuss a number of technical and regulatory options to address these concerns.

7.3 Private solutions

Private solutions to data market failures span a number of technical and qualitative solutions. In markets with low barriers to entry, consumers may vote with their money or voice to reward companies that institute good practices and punish those that don't. Companies may take proactive steps to build technical solutions that safeguard privacy while also ensuring they have access to data with little friction. These solutions, however, have a number of challenges associated with implementation.

Consumer trust is becoming increasingly important as companies ask users to share more and more often. Many private enterprises recognize that trust indirectly affects their ability to do business. As consumers become more aware of risks involved with sharing data, they may be more likely to opt-out of services. In more than a few of the examples cited thus far, public outrage was enough to illicit a swift response to concerns. Moreover, as competition increases, privacy and security quickly become features that differentiate products. There is, however, a collective action problem whereby no company has an incentive to pay costs associated with building new technological or business solutions. More importantly, new entrants face enormous network externalities when trying to convince users to switch services. Even a goliath like Google has found it difficult to lure users away from Facebook with the promise of more precise control over what data other they share.

At the same time, companies also sense a rent seeking opportunity in proactively seeking some forms of Stiglerian regulation. To this end, a number of private actors have come together at venues such as the World Economic Forum to outline frameworks and future business models that address the market failures surrounding personal data. The defining characteristic of these proposals is a shift towards a "user-centric" model. Companies that succeed in writing the definitions of these terms may

secure a competitive, first mover advantage as firms who find themselves far from new norms must pay a cost to update. These reports identify four key principles of new, user-centric data models: transparency, trust, control, and value.

Transparency mandates that users can expect to be informed about which data a company is collecting from them, who else has access, and how this data is being used. This involves simplifying privacy policies and terms of service agreements so that consumers can make informed decisions regarding what they are comfortable sharing may limit public relations problems like those Facebook confronted. Moreover, showing users what data has been collected about them provides an opportunity to correct inaccuracies and give users additional benefits. Companies whose current data policies more closely match future regulations may spare themselves from confusion or backlash from users who suddenly become aware of what old policies lacked. A company may also seek to inflict damage on competitors by pushing for requirements related data access systems that they already meet.

Trust is a more complicated issue. In a qualitative sense, trust between consumer and company is related to business fundamentals like brand-loyalty. Users may not sign up for Gmail if they don't trust Google to keep their private relationships private. "Do Not Track" settings on browsers, which prevent a user's data from being transmitted to advertisers, are one such self-regulation aimed at building this trust. In cases where informal forces are not enough, consumers may be willing to buy trust. For example, a company could offer not to collect data for the price of forgone advertising revenue. Advertising supported versions of Amazon's Kindle e-book reader may hint at this model. As advertisers increasingly track users to target them more effectively, paying extra to remove ads may amount to gaining privacy. If the company then fails to deliver on its promise, it will be subject to legal action. So long as consumers believe this threat is enough to force a company to honor its word, trust will be established. However, this economic solution makes trust and privacy available only to those who can afford to pay for it.

One alternative to the above economic solution to is a technical solution such as a trust framework. According to a 2011 World Economic Forum report, a trust

framework refers to “a set of legal and technical structures to govern the interactions of participants within the ecosystem."[194] It is agreed upon by consumers, governments, and enterprises, then implemented by some impartial authority. The report identifies two major benefits of trust frameworks. First, they limit the proliferation of personal data across many different parts of the web and second, they act as a verification system for private entities.

For example, an online movie streaming service such as Netflix would to like to ensure that customers who sign up have the intention and means of paying for it. Currently, Netflix requires users to create an account and enter some amount of personal information including a name, address, and credit card number. However, after the verification step has passed, there is no immediate reason for Netflix to retain or use an individual’s address. The data becomes another potential liability for security breach. A trust framework establishes an identity service provider as a middleman. An individual provides data to the identify service once, then any company such as Netflix can get verification of an individual’s credentials through the service. Rather than creating a copy of personal data, the identity service provider offers a simple, “verified or not" designation. Trust frameworks limit the proliferation of personal data and thus reduce the threat of privacy or safety violations while still providing a low friction way to complete transactions.

Though trust frameworks solve many problems in theory, the challenge lies in implementation. First, adopting such a system will almost always involve a company sacrificing access to data. Current trust issues between users and companies do not seem large enough to incentivize such a drastic shift of power. Second, developing and maintaining a trust framework requires the creation and operation of some third party agency. Developing such an organization will undoubtedly have costs. Standard operating procedures must be defined and refined, costing still more time and money. Meanwhile, companies have an incentive to free ride, letting others pay these costs then adopting the final product when they are completed. Finally, it is unclear how such a system would disrupt the business models of many data collectors. Removing the ability of a company to derive value from their data may change the viability of

services dramatically.

Control moves a step beyond trust. Control gives users the power to manage their own data, who can access it, and how it can be used. Current practice goes as far as to give the user detailed privacy options that dictate how different data is shared to with different entities on the web. For example, Facebook allows users to specify which groups of users can see an uploaded image and requires third party applications to acquire opt-in approval before accessing data. These measures, however, do not necessarily provide users with access to their own data. One way to provide this feature is to build upon trust frameworks and provide personal data lockers. Each individual could create a personal data locker similar to the cloud storage services available today[179]. When a user signs up for a service, all of the data they generate would be stored in the locker. The company collecting data would receive permission to access only required data from the locker or could obtain verification via a trust framework when needed. Such a system would allow individuals to access and use their own data, as well as carefully control how others use it by changing permissions. If a user wished to cancel a service, they would simply disallow that service from accessing the locker. Personal data lockers clarify property rights and permissions while limiting the ability of companies to impart negative externalities on individuals.

OpenPDS is a new technology that promises to implement this very protocol[73]. Though no applications currently use the technology, a number of major telecommunications companies have signed on as partners. It remains to be seen if business models can be found that work on these platforms, but technological solutions to problems of privacy and ownership are making leaps forward.

A final technical solution available is a form of digital rights management (DRM). DRM is a layer of permissions that is attached to and travels with each piece of data, specifying who can use a piece of information and how. Such systems have been implemented for digital music files as well as software. In theory, DRM provides assurance that data is protected from certain nefarious use regardless of how many times it is transferred or copied. However, current instances of these schemes have not realized these ideals. DRM systems aimed at preventing piracy and games and

music have been met with distain by users and have proven ineffective in deterring those serious about breaking them. However, in these cases, DRM was implemented by a few interests (i.e. record labels) to control use by many (i.e. music consumers). In the case of personal data assets, DRM would be implemented by large numbers of consumers to control use by a few companies. It may be easier to enforce digital rights in the case of personal data where companies are large and public.

The technical solutions presented by consortiums of private interests seek to clarify property rights and mitigate risks associated with personal data while still allowing room for innovation. However, many of these solutions face serious problems during implementation. In practice, all require some level of organization and enforcement on the part of governments. The remainder of this paper discusses the current and future regulatory climate surrounding personal data.

7.4 Government Regulation

Given the major challenges faced by private solutions to failures ailing data markets, government regulation is required. However, the growth of data assets has far outpaced regulations to govern them. In the United States, the few regulations that do exist were written before the rise of information giants like Google or Facebook. Though many regulatory frameworks are not yet in their final forms, significant differences can already be seen in the approaches of the United States and European Union. While the E.U. has already passed precautionary directives regulating specific practices, the U.S. appears to be leaning towards more flexible, context dependent regulation.

In 1995, the European Union adopted a Data Protection Directive to “set up a regulatory framework which seeks to strike a balance between a high level of protection for the privacy of individuals and the free movement of personal data within the European Union."[1] The directive imposed strict guidelines requiring that companies ensure data collected is accurate, was obtained with “unambiguous" consent for “legitimate" reasons, does not contain information on an certain aspects of an indi-

vidual such as religion or health, and is accessible to the user. Further, it provides individuals with the “right to object" to the use of his data and places accountability on the collectors of data to secure it. Finally, the directive provides the right to a “judicial remedy" in the cases where a user’s rights have been violated.

In early 2012, the EU released a draft of a plan to update the existing Data Protection Directive. This update seeks to introduce specific protections with respect to new services such as social networks and cloud storage. It provides users with a “right to be forgotten", requiring companies to delete personal data at a user’s request. It removes assumed consent, mandating that companies must explicitly ask users for data permissions. In addition, it frames easy access to personal data in a machine-readable format as a “right of portability". For the private sector, it attempts to reduce the administrative burden of compliance by establishing a “single national data protection authority" in each country. For consumers, it names these national authorities as the entity with which individuals can submit grievances.

The EU’s Data Protection Directive is rigid in its regulatory approach, forbidding behaviors then listing many specific exceptions. For example, it expressly forbids processing “personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and the processing of data concerning health or sex," then proceeds to outline ten specific exemptions. Moreover, the EU’s approach is not context dependent. For example, they do not allow data collectors to assume consent depending on the situation.

Facebook appears to be a popular target for testing directives. One user in Austria requested a copy of all the data Facebook had collected on him, only to receive over 1200 pages of information. As the story spread, requests overwhelmed Facebook, forcing them to provide an automated tool for downloading data[197]. Another clash occurred a few years later in early 2012 when Facebook tried to roll out a new facial recognition feature. When a user uploaded photos, they were given the choice of allowing Facebook to perform image analysis to identify and label individuals in photos without those individuals’ permission. The Irish Data Protection Commissioner (Facebook’s European operations are based in Ireland) issued an assessment suggest-

ing that implementing this feature would violate the Data Protection Directive. This assessment and public backlash caused Facebook to delay the feature pending further iteration.

Google has also run into problems in Europe. Some countries, such as Germany and Austria, have forbade Google from compiling photographs of public roads data for use in their Street View service without at least providing a way for citizens to opt out. More recently, it was revealed that in the process of collecting images and data for location-based services, a software glitch inadvertently intercepted data being sent across open Wi-Fi networks[2]. A number of EU nations responded by demanding Google reveal which data had been obtained and threatened to levy steep fines. Though fines were never issued in the EU, the cost of compliance may have factored into Google's decision to abandon efforts to expand Street View coverage in some areas[149]. Though the updates to the data protection directive list regulatory consolidation as a goal, companies must still negotiate with each state's national authority individually. In general, the regulatory approach of the EU has broadened the property rights of consumers. The response to these regulations, by U.S. based companies at least, has been to delay or abandon investment in new technologies.

The regulatory strategy of the United States even more complex and disaggregate. According to the Obama Administration's white paper on "Consumer Data Privacy in a Networked World," "...Federal data privacy statutes apply only to specific sectors such as healthcare, education, communications, and financial services, or, in the case of online data collection, to children." For example, the Health Insurance Portability and Accountability Act (HIPAA) regulates the collection and use of personal health information. It restricts entities such as insurance providers from sharing health records or, in some cases, using records to determine services and premiums. It also gives individuals the right to access and correct health information. The Fair Credit Reporting Act (FCRA) outlines similar guidelines for data relating to a consumer's credit information. Finally, the Children's Online Privacy Protection Act (COPPA) requires online entities to obtain parental permission before providing services to children under the age of 13.

While the United States Department of Health and Human Services (HHS) Office enforces HIPAA for Civil Rights, the FCRA and COPPA fall under the jurisdiction of the Federal Trade Commission (FTC). In addition, the FTC has adopted Fair Information Processing Standards (FIPS) that provide guidelines concerning use of electronic consumer data. These guidelines, however, are not enforceable. Moreover, they were established in 1995, before Google even existed. Finally, the Federal Communications Commission (FCC) is responsible for enforcing regulations relating to communications data sent over the Internet or phone networks.

Moving forward, the Obama administration has signaled a desire to introduce comprehensive consumer data privacy regulations enforceable by the FTC. For its part, the FTC has acknowledged it has already established FIPs guidelines and is willing and capable of taking on this role. The Obama administration has urged congress to draft and pass legislation creating a “Consumer Privacy Bill of Rights" to “give consumers clear guidance on what they should expect from those who handle their personal information, and set expectations for companies that use personal data."[3] The Consumer Privacy Bill of Rights provides for many of the concepts addressed earlier. Consumers should expect control, transparency, accuracy, and security with regards to personal data. However, there are also calls for more “focused collection" of data, whereby companies gather and store as little information as possible. Finally, the administration notes that the regulations added would not affect the enforcement of existing frameworks such as HIPAA.

To accompany the administration’s proposal, the FTC released a report in 2012 titled “Protecting Consumer Privacy in an Era of Rapid Change." This report summarized the regulatory strategy advocated by the FTC in the event Congress grants it enforcement powers. These principles reflect the policy wishes of the Obama administration as well as nearly 450 public comments from industry and consumer groups. The distinguishing characteristic between the desired approach of the U.S. and the approach taken by the EU is a “respect for context." Rather than outline a set of specifically forbidden actions, the U.S. wishes to allow companies to “implement the privacy protections of the framework in a way that is proportional to the nature,

sensitivity, and amount of data as well as the size of the business at issue." Privacy standards will depend on a "context of interaction" such that "the need for choice [to opt-in or out of data permission] should depend on reasonable consumer expectation." For example, the FTC would not limit how a company could use data it collected for internal marketing purposes.

The FTC report also focuses heavily on self-regulation. It highlights initiatives like that of the Digital Advertising Alliance (DAA), a consortium of companies responsible for nearly 90% of all online marketing, which developed an icon labeling tool to educate consumers on where targeted ads appear along with committing to honor "Do Not Track" standards. Finally, the FTC and Obama administration push for "privacy by design" whereby companies design new products and services with privacy as a specific design principal integrated into specifications from the start.

Though the EU and U.S. differ somewhat on regulatory philosophy, both recognize the importance of interoperability across nations. The Internet often blurs geopolitical boundaries. Data servers need not be located in countries where their owners operate. Moreover, one of the greatest features of the Internet is its accessibility from anywhere. Forcing companies to create different data architectures and practices depending on where a user is connecting from may prove enough a burden to stifle innovation (see Google Street View in Germany). As part of the 1995 EU Data Protection Directive, European companies could send data outside the union if privacy protection laws in the other country or company meet the standards of the EU. This safe harbor provision has allowed U.S. companies to operate in the E.U. and vice versa.

In general, the regulatory frameworks of the U.S. and E.U. both focus on permissions as opposed to ownership rights. Both policies require companies to provide more transparent terms of service to assist the consumer's ability to provide informed consent. Though the U.S. leaves room for companies to assume consent in some instances, both EU and U.S. policy provide the right of the user to opt-out at any time. In general, both regulatory platforms allow first parties (data collectors) to use data freely for internal purposes with exceptions for certain cases including health, credit, or children. Both policies require first parties acquire explicit informed consent

from users before offering personal data to third parties. This requirement, in theory, provides rights even to users who have not provided data to a company, but whose privacy may be at risk if their friends have. The majority of these guidelines would integrate well with technical solutions such as DRM or personal data lockers.

Requiring companies to disclose to users what data they have collected and what they are using it for may help users determine when their safety or privacy is at risk, allowing them to opt out. However, the biggest impact on the potential negative externalities associated with personal data comes from the creation of clear authorities to enforce policy. In the U.S., for example, complaints of poor data practices could be made to the FTC who would have the power to investigate and punish companies, like they did in the case of Google regarding the accidental collection of Wi-Fi data. The threat of fines or other punishment may help companies internalize the costs of security breaches.

It is more difficult to predict how these policies may affect companies. On one hand, the more rigid regulatory statutes in the EU may act as a burden that stifles innovation. On the other hand, the more flexible, self-regulatory approach favored by the U.S. seems more flexible for a rapidly progressing industry, but may also be vulnerable to Stiglerian rent grabs. Incumbent companies have large incentives to promote policies and standards that protect their business models while barring entrants.

7.5 Conclusions

The exponential increase in personal data collected and stored online has created a number of challenges for consumers, private enterprises, and public institutions. With huge economic potential, balancing interests in a way that promotes growth and innovation is critical. Three market failures were identified, uncertain property rights, information deficits and asymmetries, and externalities. Uncertainties related to who can use and sell data are increasingly leading to substantial tort actions imposing real costs to companies and users. Information deficits and asymmetries may lead to

suboptimal market outcomes where data is underutilized or over shared. Negative externalities threaten the privacy and security of individuals while positive externalities may provide real benefits by solving problems related to public transportation or health care.

To address these market challenges, a number of technical solutions have been proposed by private enterprise. They include the establishment of trust frameworks and data lockers, allowing companies to verify necessary identities and access data without requiring a user to create another copy or give up control. Other companies have voluntarily introduced “Do Not Track" features to curtail what data they collect. Finally, digital rights management would allow rights to travel with data, reducing risks associated with data proliferation to third parties. However, serious challenges exist in the implementation and coordination of these solutions. It is clear that regulation will play a large role in establishing and enforcing standards and guidelines.

The goals of regulations, independent of the approach, are three fold. First, they require firms to state property rights in a way that can be easily understood by consumers. In some cases, they explicitly define rights. Second, they seek this transparency to reduce information asymmetries that lead to over sharing. Third, they seek to limit the potential damages caused by negative externalities in the case of security and privacy. While the EU touts decreased administrative complexity as a benefit to companies, they do not go nearly as far as US proposals that explicitly state the goal of ensuring that data continues to flow to create economic value. These regulatory frameworks are far from static. There is no doubt they will be updated and refined as new issues arise.

Moving forward, the most significant next step regarding the personal data will be how regulation is implemented in the United States. Until then, industry seems content to collect and use as much data as possible and users are willing to give it to them. In certain instances, usually involving very public players, consumers may find some power in public outrage that thus far has been loud enough to force action, but as the economic incentives grow, individuals may see their position weakened. Ultimately, there are tremendous opportunities to provide value for companies and

improve the lives of billions by leveraging personal data to provide new services public and private. Any action taken by industry or governments should seek to protect individuals while allowing data to flow with as little friction as possible.

Chapter 8

Conclusion

As millions of people migrate cities each year, it is critical that we put new data sources to work improving them. The density of cities brings economic productivity, provides cultural amenities, and facilitates sustainability, but is also the root of problems related to congestion, health, and safety. This thesis has presented new perspectives and applications of massive, passively collected data sources to understand human mobility and its relation to these outcomes. It has detailed a number of new methods and applications for extracting meaningful insights from this data to make cities better places to live. Above all, though, this work highlights the need for interdisciplinary approaches. We have seen combinations of methods from data mining and machine learning, network theory, and spatial statistics, with deep domain knowledge from urban planning, economics, and transportation. For decades, planners and sociologists such as Jane Jacobs have recognized cities as systems of organized complexity where the interactions between millions of residents combine to produce emergent phenomena. The promise of "big data" is to measure these interactions and enable us to study their consequences at a massive scale.

This thesis builds on the maturing set of works that quantify and model human mobility patterns using data collected from mobile phones and other digital resources. With penetration rates over 100% in much of the developed world and over 80% in parts still developing, these devices capture high resolution spatiotemporal and social data on millions of individuals within a city. From this, we have learned how to ex-

tract patterns of human movement and social interactions at a scale unachievable by traditional survey or census. In addition to high resolution behavioral data, extensive geographic data on transportation and city infrastructure is now available to anyone via crowd- and open-sourced databases. The work presented here furthers these goals by developing and applying novel methods to uncover and model the interaction between city infrastructure, demand for it, and its role in social and economic behaviors.

Perhaps the most promising aspect of this new approach to understanding urban systems is its scalability. The ubiquity digital devices generate data in similar formats and volumes across the world, allowing methods to be applied more broadly than ever before. While it is critical that methods be validated against the best available ground truth data, there is now little cost to generating results for additional cities. For example, in the case of travel demand estimation, we have shown that the same model may be applied to five cities, producing consistent results across each. To do similar work via traditional data collection methods would have taken decades and would come at a cost greater than many cities or nations could afford. These new data sources provide fast, accurate, and inexpensive access to critically important planning tools.

Throughout this thesis, we saw these broad interdisciplinary and generalizable approaches applied across three broad domains: social behavior and choice, economic outcomes, and mobility and city infrastructure infrastructure.

8.1 Social Behavior and Choice

At their core, mobile phones are communications technologies allowing individuals exchange ideas and information with social contacts. Data from the social networks generated by use of these devices and the applications that run on them has been analyzed extensively over the past decade, revealing much about how we build and maintain relationships. These social interactions wield great influence the choices of individuals. To ensure models of demand account for these forces, Chapter 3 details

methods for understanding how social behavior relates to related to mobility. While data scarcity has generally required mobility and social behavior to be treated as independent, more recent work has made use of huge, location based social networks to quantify the close relationship between the two.

While many of these works focus on the distance between two social contacts, far less is known about similarities in the places two individuals choose to visit within a city. This work quantified social and mobility similarity within the cities and created methods which use mobility information to add context to the nature of social relationships.

We saw that two individuals who call each other are far more similar than random strangers and mobility similarity is positively correlated with tie strength in social networks. Individuals who tend to be more social are also visit more distinct locations. This similarity makes it possible to reconstruct large amounts of an individual's visitation patterns from the visitations of their social contacts. The gains from combining social and movement data go both ways. By examining when during the day two individuals visit the same places, it is possible to classify the nature of a social relationship. This result may help add much needed context to large social networks. Finally, we saw that a extension of an established mobility model to include visitation preferences based on the choices of social contacts was able and required to replicate empirical results.

Social influence is a strong motivator. It influences decisions and choices people make on where to go, what to buy, and what to believe. If left out of our models, a large part of our decision making process will be lost, decreasing their accuracy. Moreover, quantifying the level of similarity between social contacts is becoming increasingly important to building new transportation solutions. Car and ride sharing services like Uber and Lyft are aggressively rolling out new services that use real-time algorithms to match riders going to and from similar places. These services have the potential to dramatically reduce both the direct and indirect costs of travel, lowering congestion and the environmental impact of it. While new technologies and the data they collect making these services possible, large issues involving trust and reputation

remain. The work presented in this thesis suggests social networks are a good place to start. Not only are social contacts already similar, making matching easier, but there is also presumably some level of trust and comfort already established, helping with adoptions. Combining these results with the new services made possible by smart phone adoption will enable the next generation of social applications to promote real change in people's behaviors their impact on cities and the environment.

8.2 Economic Behavior in Cities

The ultimate goal of understanding and improving city infrastructure that it may positively impact the lives of people. To measure and model these benefits more explicitly, Chapter 4 applied domain knowledge of human mobility to efforts in computational social science with a focus on urban economics. At the micro level, we showed that it is possible to detect a mass layoff in a town and identify affected individuals by observing changes in calling behavior. Moreover, we found that mobile phones are exceptional sensors for measuring the impact of layoffs on individuals, showing persistent drops in a user's social and mobility behavior. We then leveraged the size of this data to scale our study from the micro- to macro-level, a rare opportunity in economics. By incorporating measured changes in social and mobility behavior in populations of users into predictions of unemployment rates at the regional level, forecasts of unemployment rates can be made and improved months before traditional surveys are conducted and released.

The work in this chapter has a number of far reaching consequences. Urban economists and transportation planners have long understood that mobility and social networks provide access to new and better opportunities for employment. The sharp decline in both following a mass layoff hints and why these events have such lasting negative consequences for those affected. These results also uncover social behavior and mobility has potential levers for helping the unemployed recover. While current benefits typically focus on monetary aspects of unemployment, this work shows that there are other areas that require attention as well.

More broadly the social sciences have a rich literature documenting the impact of spatial variables on outcomes. We are well aware, for example, that cities are segregated based on demographic and economic characteristics and that this segregation is often correlated with diminished opportunity and economic outcomes for less fortunate individuals. Accessibility and mobility allow individuals to find better jobs with shorter commutes, and live in areas where they can freely mix and exchange ideas with others. These benefits are not afforded everyone. With the ability to measure interactions between people and infrastructure, there has never been a better time to explore these relationships further and strengthen the link between urban planning and outcomes for residents. Doing so, however, requires more interdisciplinary approaches that can transform new data resources into sound conclusions and applications.

8.3 Mobility and City Infrastructure

Understanding demand for and operation of city infrastructure is crucial for civil engineers, planner, and policy makers. Passively collected data from new mobile devices have generated a wealth of insights about human movement. Metrics and methods from statistical physics have been applied to data on millions of individuals to show that we are generally slow to explore new places, highly predictable, but also unique. This data has been used in conjunction with traditional survey and census sources to produce and validate estimates of aggregate flows of people or vehicles from origins and destinations around the city. Chatpers 5 and 6 presented mapping these mobility insights and measurements onto transportation infrastructure and zoning information in order to evaluate system performance.

Chapter 5 describes a software system that uses call detail records from millions of mobile phones to estimate travel demand and road usage patterns in multiple cities around the world. The system takes as an input billions of rows of data on where mobile phone users make calls or send messages from and estimates the flow of people between origins and destinations in a city at different times and for different purposes.

These origin-destination matrices are then used to assign trips to transportation infrastructure. For example, we leverage crowd-sourced road networks from repositories such as OpenStreetMap and routes vehicle trips through road networks. The result is a rich map of traffic and congestion which contains estimated usage data for nearly every road in the city. Because the system tracks the origins and destinations of every vehicle routed, we are also able to perform a detailed analysis of which areas contribute traffic to a particular congested area or which streets serve local traffic versus traffic from all over a city. Finally, we have seen an interactive, online visualization platform that allows researchers and policy makers to explore travel and congestion patterns of cities.

This system was built with generalizability as a requirement. It has currently processed more multiple cities around the world and new cities can be added easily provided input data. We have gone to great lengths to validate this work against traditional methods that use survey and census data and compare with the state of the art in the transportation and urban planning communities. This platform has great potential to produce estimates of travel demand more accurately, faster, and at lower cost than traditional survey based methods. The relative uniformity of mobile phone data across the world presents a great opportunity to conduct comparative studies of congestion patterns in cities and explore the relationship between network topology, system performance, and travel demand. Still more can be done to include multiple modes of transportation such as busses, rail, and even bicycles. Finally, there is ability to connect this work with that from the previous chapters in this thesis to understand how social and economic behaviors are correlated with transportation network usage. This work is critical into leveraging the massive scale of new mobility data to understand and improve the transportation systems we use every day.

In addition to transportation infrastructure, land use and zoning regulations are used to guide the form and growth of cities. These regulations are largely static and do little to suggest how an area might be used over the course of a day or week. Mobile phone activity within an area was explored as a proxy for the zoning regulations in that same area. Chapter **??** showed that this activity was a poor predictor of

zoning classifications. However, a deep analysis of classification errors reveals strong temporal patterns in mobile phone activity that suggest that classifications may be incorrect. The addition of high resolution data on the location and type of points of interest in an area does little to improve accuracy and strengthens this conclusion. These results highlight the ability of new data sources to complement and improve the old. They are dynamic and quickly reflect changes in behavior allowing policy makers to more easily understand and monitor the form and function of cities.

8.4 Future Directions

While all of these applications of new data sources are promising contributions, it is important to remember that this data was collected from people. Moving forward, very real concerns over the use of this data must be addressed. Chapter 7 seized upon the definition of this data as an economic asset and discussed implications of its use from this perspective. Markets for this non-rival good are plagued with inefficiencies due to uncertain property rights, large externalities, and information asymmetries. Governments have taken divergent approaches to the regulation of personal data and these discussions must continue as technology advances. Finding a balance between privacy and safety and allowing private entities to create leverage data to create valuable resources is imperative.

The potential of these new data sources will only grow as mobile phone penetration and application adoption rates increase. The these devices will become more accurate and collect more data. Data storage and computation systems will grow in scale and power and analytics algorithms will continue to improve what can be measured and predicted. This growth alone is not enough to realize this data's potential. We must be more conscious of its biases and limitations. We must respect the rich foundation of knowledge gained from previous work making use of carefully curated "small data". Rather than blindly applying algorithms, more care must be taken to consider the processes that generate these data rather than it's pure predictive power. With those considerations in mind, there are a number of areas related to human mobility and

behavior within urban systems can benefit from additional data-driven approaches.

Although progress has been made to transform new or traditional data sources into usable travel demand estimates, more can be done to understand how to translate between the two. Promising work seeks to administer traditional travel surveys via mobile devices that collect data in new formats. Translating and correlating between the self-reported survey data and the digital breadcrumbs collected by device use will give a far more complete picture of the limitations and promise of each.

In addition to a better understanding of this data, there are a number of strategic areas that could benefit from data-drive approaches. Just as this thesis has presented an implementation of a four step model to create origin-destination matrices from new data sources, the assignment of those flows to transportation networks remains an area ripe for disruption. As GPS data is increasingly collected via mobile devices, we are now able to collect high resolution data on the exact paths that individuals take to navigate cities. Applying data mining techniques may shed new light on the internal models and heuristics that govern route and mode choice. These insights are critical to designing better control strategies for smart infrastructure.

Above all, however, the most value to be gained from these data is layering them on top of one another. Data integration makes it possible to add context to billions of anonymized call detail records and billions of these records make it possible to measure better estimate travel demand. The two must be used together in order to unlock the potential of each. Moving forward, more complex layerings of data must be explored. For example, by incorporating demographic information from census tracts into the OD matrices estimated from billions of phone calls, it becomes possible to explore not only long studied patterns of static segregation of cities, but how that segregation affects where individuals move dynamically throughout the day.

We are now able to quantify the organized complexity of cities and measure the daily interactions between people and places. This ability comes at an important time when understanding is needed more than ever. As the next billion people stream into cities all over the world, leveraging this data in responsible ways will help ensure that cities remain places of opportunity for their residents.

Bibliography

[1] European parliament and council directive 95/46/EC of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, October 1995.

[2] WiFi data collection: An update, May 2010.

[3] Consumer data privacy in a networked world: A framework for protecting privacy and promoting innovation in the global digital economy. Technical report, White House, Washington, DC, February 2012.

[4] Statement of rights and responsibilities, January 2015.

[5] Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. Social fMRI: Investigating and shaping social mechanisms in the real world. In *Pervasive and Mobile Computing*, volume 7, pages 643–659, 2011.

[6] Rahmi Akcelik. Travel time functions for transport planning purposes: Davidson's function, its time dependent form and alternative travel time function. *Australian Road Research*, 21(3), 1991.

[7] Lauren P. Alexander, Shan Jiang, Mikel Murga, and Marta C González. Validation of origin-destination trips by purpose and time of day inferred from mobile phone data. 2014.

[8] Dolan Antenucci, Michael Cafarella, Margaret Levenstein, Christopher Ré, and Matthew D Shapiro. Using Social Media to Measure Labor Market Flows. Technical report, National Bureau of Economic Research, March 2014.

[9] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: detecting influenza epidemics using Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1568–1576. Association for Computational Linguistics, 2011.

[10] Yasuo Asakura and Eiji Hato. Tracking survey for individual travel behaviour using mobile communication instruments. *Transportation Research Part C: Emerging Technologies*, 12(3):273–291, 2004.

[11] Nikos Askitas and Klaus F Zimmermann. Google econometrics and unemployment forecasting. 2009.

[12] Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pages 61–70, 2010.

[13] James P. Bagrow and Yu-Ru Lin. Mesoscopic structure and social aspects of human mobility. *PloS one*, 7(5):e37676, January 2012.

[14] James P Bagrow, Dashun Wang, and Albert-László Barabási. Collective Response of Human Populations to Large-Scale Emergencies. *PLoS ONE*, 6(3):8, 2011.

[15] Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J Ramasco, and Alessandro Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51):21484–9, December 2009.

[16] Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J Ramasco, and Alessandro Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51):21484–9, December 2009.

[17] D. Banister. Reducing the need to travel. *Environment and Planning B: Planning and Design*, 24(3):437–449, 1997.

[18] Hillel Bar-Gera. Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from israel. *Transportation Research Part C: Emerging Technologies*, 15(6):380–391, 2007.

[19] Albert-László Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

[20] Marc Barthélemy. Spatial networks, 2011.

[21] Holger Bast, Stefan Funke, Peter Sanders, and Dominik Schultes. Fast routing in road networks with transit nodes. *Science*, 316(5824):566–566, 2007.

[22] Michael Batty. The size, scale, and shape of cities. *Science (New York, N.Y.)*, 319(5864):769–771, 2008.

[23] Vitaly Belik, Theo Geisel, and Dirk Brockmann. Natural human mobility patterns and spatial spread of infectious diseases. *Physical Review X*, 1(1):011001, 2011.

[24] Michael GH Bell. The estimation of origin-destination matrices by constrained generalised least squares. *Transportation Research Part B: Methodological*, 25(1):13–22, 1991.

[25] Moshe E. Ben-Akiva and Steven R. Lerman. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press, 1985.

[26] R Benabou. Workings of a city: location, education, and production. *The Quarterly Journal of Economics*, 108:619–652, 1993.

[27] Luís M A Bettencourt. The origins of scaling in cities. *Science*, 340(6139):1438–41, June 2013.

[28] Luís M A Bettencourt, José Lobo, Dirk Helbing, Christian Kühnert, and Geoffrey B West. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences of the United States of America*, 104(17):7301–7306, 2007.

[29] Luís M A Bettencourt and Geoffrey West. A unified theory of urban living. *Nature*, 467:912–913, 2010.

[30] Nick Bilton and Brian Stelter. Sony says PlayStation hacker got personal data. *The New York Times*, April 2011.

[31] Joshua Evan Blumenstock, Marcel Fafchamps, and Nathan Eagle. Risk and Reciprocity Over the Mobile Phone Network: Evidence from Rwanda. *SSRN Electronic Journal*, November 2011.

[32] Marián Boguñá and Romualdo Pastor-Satorras. Class of correlated random networks with hidden variables, 2003.

[33] Douglas Bownman. Goodbye, google, March 2009.

[34] David Branston. Link capacity functions: A review. *Transportation Research*, 10(4):223–236, 1976.

[35] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001.

[36] D Brockmann, L Hufnagel, and T Geisel. The scaling laws of human travel. *Nature*, 439:462–465, 2006.

[37] Dirk Brockmann, Lars Hufnagel, and Theo Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[38] Michael C Burda, Daniel S Hamermesh, and Jay Stewart. Cyclical Variation in Labor Hours and Productivity Using the ATUS. Technical report, National Bureau of Economic Research, December 2012.

[39] N Caceres, JP Wideberg, and FG Benitez. Deriving origin destination data from a mobile phone network. *Intelligent Transport Systems, IET*, 1(1):15–26, 2007.

[40] Francesco Calabrese, Massimo Colonna, Piero Lovisolo, Dario Parata, and Carlo Ratti. Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome, 2011.

[41] Francesco Calabrese, Mi Diao, Giusy Di Lorenzo, Joseph Ferreira Jr, and Carlo Ratti. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation research part C: emerging technologies*, 26:301–313, 2013.

[42] Francesco Calabrese and Carlo Ratti. Real Time Rome. *NETCOM*, 20(3-4):247–258, 2006.

[43] Francesco Calabrese, Jonathan Reades, and Carlo Ratti. Eigenplaces: Segmenting Space through Digital Signatures. *IEEE Pervasive Computing*, 9(1):78–84, January 2010.

[44] Julián Candia, Marta C González, Pu Wang, Timothy Schoenharl, Greg Madey, and Albert-László Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015, 2008.

[45] David Card. Origins of the Unemployment Rate: The Lasting Legacy of Measurement without Theory. *American Economic Review*, 101(3):552–557, May 2011.

[46] Ennio Cascetta. Estimation of trip matrices from traffic counts and survey data: a generalized least squares estimator. *Transportation Research Part B: Methodological*, 18(4):289–299, 1984.

[47] Serdar Çolak, Lauren P. Alexander, Bernardo G. Alvim, Shomik R. Mehndiretta, and Marta C. González. Analyzing cell phone location data for urban travel: Current methods, limitations, and opportunities. *to appear in TRB Proceedings*, 2015.

[48] Damon Centola. An experimental study of homophily in the adoption of health behavior. *Science*, 334(6060):1269–1272, 2011.

[49] Robert Cervero and Kara Kockelman. Travel demand and the 3Ds: Density, diversity, and design. *Transportation Research Part D: Transport and Environment*, 2(3):199–219, September 1997.

[50] Kathy H. Chan. On facebook, people own and control their information, February 2009.

[51] Sewin Chan and Ann Huff Stevens. Employment and retirement following a late-career job loss. *American Economic Review*, pages 211–216, 1999.

[52] Kerwin Kofi Charles and Melvin Stephens Jr. Job displacement, disability, and divorce. Technical report, National bureau of economic research, 2001.

[53] Y. Chen and S. Rosenthal. Local amenities and life-cycle migration: Do people move for jobs or fun? *Journal of Urban Economics*, 64(3):519–537, November 2008.

[54] Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Z. Sui. Exploring Millions of Footprints in Location Sharing Services. In *ICWSM*, pages 81–88, 2011.

[55] Raj Chetty, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States. January 2014.

[56] Cynthia Chew and Gunther Eysenbach. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. *PloS one*, 5(11):e14118, 2010.

[57] Albert M. L. (Albert Man Loon) Ching. A user-flocksourced bus intelligence system for Dhaka, 2012.

[58] Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining KDD 11*, KDD '11, page 1082. ACM Press, 2011.

[59] Hyunyoung Choi and Hal Varian. Predicting initial claims for unemployment benefits. *Google Inc*, 2009.

[60] Hyunyoung Choi and Hal Varian. Predicting the Present with Google Trends. *Economic Record*, 88:2–9, 2012.

[61] Nicholas A Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *The New England journal of medicine*, 357(4):370–379, 2007.

[62] Timothy J Classen and Richard A Dunn. The effect of job loss and unemployment duration on suicide risk in the United States: a new look using mass-layoffs and unemployment duration. *Health economics*, 21(3):338–350, 2012.

[63] Serdar Colak, Christian M Schneider, Pu Wang, and Marta C González. On the role of spatial dynamics and topology on network flows. *New Journal of Physics*, 15(11):113037, 2013.

[64] Vittoria Colizza, Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences of the United States of America*, 103(7):2015–20, February 2006.

[65] C.D.a Cottrill, F.C.a Pereira, F.a Zhao, I.F.b Dias, H.B.c Lim, M.E.d Ben-Akiva, and P.C.d Zegras. Future mobility survey. *Transportation Research Record*, (2354):59–67, 2013.

[66] David J Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences of the United States of America*, 107(52):22436–22441, 2010.

[67] R. Crane. The Influence of Urban Form on Travel: An Interpretive Review, 2000.

[68] Aron Culotta. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the first workshop on social media analytics*, pages 115–122. ACM, 2010.

[69] Carlos F Daganzo. Optimal sampling strategies for statistical models with discrete dependent variables. *Transportation Science*, 14(4):324–345, 1980.

[70] Damon Darlin. Google settles suit over buzz and privacy, November 2010.

[71] Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the Crowd: The privacy bounds of human mobility. *Nature Scientific Reports*, 3:1376, 2013.

[72] Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3, 2013.

[73] Yves Alexandre De Montjoye, Erez Shmueli, Samuel S. Wang, and Alex Sandy Pentland. OpenPDS: Protecting the privacy of metadata through SafeAnswers. *PLoS ONE*, 9, 2014.

[74] Yves-Alexandre de Montjoye, Zbigniew Smoreda, Romain Trinquart, Cezary Ziemlicki, and Vincent D. Blondel. D4D-Senegal: The Second Mobile Phone Data for Development Challenge. July 2014.

[75] Ed de Quincey and Patty Kostkova. Early warning and outbreak detection using social networking websites: The potential of twitter. In *Electronic healthcare*, pages 21–24. Springer, 2010.

[76] P. Deville, C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E. Gaughan, V. D. Blondel, and A. J. Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, pages 1408439111–, October 2014.

[77] Giusy Di lorenzo, Marco Luca Sbodio, Francesco Calabrese, Michele Berlingerio, Rahul Nair, and Fabio Pinelli. AllAboard. In *Proceedings of the 19th international conference on Intelligent User Interfaces - IUI '14*, pages 335–340, New York, New York, USA, February 2014. ACM Press.

[78] Manlio De Domenico. Interdependence and Predictability of Human Mobility and Social Interactions. *csbhamacuk*, 2012, 2012.

[79] Nathan Eagle, Michael Macy, and Rob Claxton. Network diversity and economic development. *Science (New York, N.Y.)*, 328(5981):1029–1031, 2010.

[80] Nathan Eagle and Alex Pentland. Reality mining: Sensing complex social systems. *Personal and Ubiquitous Computing*, 10:255–268, 2006.

[81] Nathan Eagle and Alex Sandy Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63:1057–1066, 2009.

[82] Nathan Eagle, Alex Sandy Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America*, 106:15274–15278, 2009.

[83] Paul S Earle, Daniel C Bowden, and Michelle Guy. Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics*, 54(6), 2012.

[84] Rochelle M Edge, Refet S Gürkaynak, RICARDO REIS, and CHRISTOPHER A SIMS. How useful are estimated dsge model forecasts for central bankers?[with comments and discussion]. *Brookings Papers on Economic Activity*, pages 209–259, 2010.

[85] Glenn Ellison, Edward L. Glaeser, and William R. Kerr. What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns. *American Economic Review*, 100(3):1195–1213, June 2010.

[86] Michael Ettredge, John Gerdes, and Gilbert Karuga. Using web-based search data to predict macroeconomic statistics. *Communications of the ACM*, 48(11):87–92, 2005.

[87] Joseph Ferreira, Mi Diao, Yi Zhu, Weifeng Li, and Shan Jiang. Information infrastructure for research collaboration in land use, transportation, and environmental planning. *Transportation Research Record: Journal of the Transportation Research Board*, 2183(1):85–93, 2010.

[88] Matthew L. Freedman. Job hopping, earnings dynamics, and industrial agglomeration in the software publishing industry. *Journal of Urban Economics*, 64:590–600, 2008.

[89] M. Gelman, S. Kariv, M. D. Shapiro, D. Silverman, and S. Tadelis. Harnessing naturally occurring data to measure the response of spending to income. *Science*, 345(6193):212–215, July 2014.

[90] Karst T. Geurs and Bert van Wee. Accessibility evaluation of land-use and transport strategies: review and research directions. *Journal of Transport Geography*, 12(2):127–140, June 2004.

[91] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2008.

[92] Edward Glaeser. *Triumph of the city: How our greatest invention makes US richer, smarter, greener, healthier and happier*. Pan Macmillan, 2011.

[93] Edward L Glaeser, Matthew E Kahn, and Jordan Rappaport. Why do the poor live in cities? the role of public transportation. *Journal of urban Economics*, 63(1):1–24, 2008.

[94] Sharad Goel, Jake M Hofman, Sébastien Lahaie, David M Pennock, and Duncan J Watts. Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences*, 107(41):17486–17490, 2010.

[95] Janaína Gomide, Adriano Veloso, Wagner Meira Jr, Virgílio Almeida, Fabrício Benevenuto, Fernanda Ferraz, and Mauro Teixeira. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. In *Proceedings of the 3rd International Web Science Conference*, page 3. ACM, 2011.

[96] Bruno Goncalves and Jose J Ramasco. Human dynamics revealed through Web analytics. *Physical Review E*, 78(2):7, 2008.

[97] Marta C González, César A Hidalgo, and Albert-László Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[98] Marta C González, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[99] Przemyslaw A. Grabowicz, Jose J. Ramasco, Bruno Goncalves, and Victor M. Eguiluz. Entangling mobility and interactions in social media. page 16, July 2013.

[100] Dino Grandoni. You may have been a lab rat in a huge facebook experiment, June 2014.

[101] Tami Gurley and Donald Bruce. The effects of car access on employment outcomes for welfare recipients. *Journal of Urban Economics*, 58:250–272, 2005.

[102] Michelle Guy, Paul Earle, Chris Ostrum, Kenny Gruchalla, and Scott Horvath. Integration and dissemination of citizen reported and seismically derived earthquake information via social network technologies. In *Advances in intelligent data analysis IX*, pages 42–53. Springer, 2010.

[103] Mark Hachman. TomTom tracks u.k. drivers for insurance purposes, February 2012.

[104] Randolph W. Hall, editor. *Handbook of Transportation Science*, volume 23 of *International Series in Operations Research & Management Science*. Springer US, Boston, MA, 1999.

[105] Xiao-Pu Han, Qiang Hao, Bing-Hong Wang, and Tao Zhou. Origin of the scaling law in human mobility: Hierarchy of traffic systems. *Physical Review E*, 83(3):2–6, 2011.

[106] Walter G. Hansen. How Accessibility Shapes Land Use. *Journal of the American Institute of Planners*, 25(2):73–76, May 1959.

[107] Jerry Hausman and Ephraim Leibtag. CPI Bias from Supercenters: Does the BLS Know that Wal-Mart Exists? Technical report, National Bureau of Economic Research, August 2004.

[108] Bernd Hayo and Ali M Kutan. The impact of news, oil prices, and global market developments on Russian financial markets1. *Economics of Transition*, 13(2):373–393, 2005.

[109] Martin L Hazelton. Estimation of origin–destination matrices from link flows on uncongested networks. *Transportation Research Part B: Methodological*, 34(7):549–566, 2000.

[110] Martin L Hazelton. Inference for origin–destination matrices: estimation, prediction and reconstruction. *Transportation Research Part B: Methodological*, 35(7):667–676, 2001.

[111] Martin L Hazelton. Some comments on origin–destination matrix estimation. *Transportation Research Part A: Policy and Practice*, 37(10):811–822, 2003.

[112] Miguel Helft. Critics say google invades privacy with new service. *The New York Times*, February 2010.

[113] J Vernon Henderson, Adam Storeygard, and David N Weil. Measuring Economic Growth from Outerspace. *The American Economic Review*, 102(2):994–1028, May 2012.

[114] Juan C. Herrera, Daniel B. Work, Ryan Herring, X. Ban, Quinn Jacobson, and Alexandre M. Bayen. Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment. *Transportation Research Part C: Emerging Technologies*, 18:568–583, 2010.

[115] C Herrera-Yagüe, C M Schneider, Z Smoreda, T Couronné, P J Zufiria, and M C González. The elliptic model for communication fluxes. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(4):P04022, April 2014.

[116] Ryan Herring, Tania Abou Nasr, Amin Abdel Khalek, and Alexandre Bayen. Using Mobile Phones to Forecast Arterial Traffic through Statistical Learning. *Electrical Engineering*, 59:1–22, 2010.

[117] J. D. Hunt, D. S. Kriger, and E. J. Miller. Current operational urban land?use?transport modelling frameworks: A review, 2005.

[118] Md Shahadat Iqbal, Charisma F Choudhury, Pu Wang, and Marta C González. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40:63–74, 2014.

[119] Md. Shahadat Iqbal, Charisma F. Choudhury, Pu Wang, and Marta C. González. Development of originâĂŞdestination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40:63–74, March 2014.

[120] Jane Jacobs. *The Death and Life of Great American Cities*, volume 71. 1961.

[121] Jerald Jariyasunant. *Improving Traveler Information and Collecting Behavior Data with Smartphones*. PhD thesis, 2012.

[122] Shan Jiang, Gaston A Fiore, Yingxiang Yang, Joseph Ferreira Jr, Emilio Frazzoli, and Marta C González. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, page 2. ACM, 2013.

[123] Stephen R. G. Jones and W. Craig Riddell. The Measurement of Unemployment: An Empirical Approach. *Econometrica*, 67(1):147–162, 1999.

[124] H.S. Kim. QoS provisioning in cellular networks based on mobility prediction techniques. *IEEE Communications Magazine*, 41(1):86–92, January 2003.

[125] Sunwoong Kim. Labor Specialization and the Extent of the Market, 1989.

[126] Eleni Kosta, Hans Graux, and Jos Dumortier. Collection and Storage of Personal Data: A Critical View on Current Practices in the Transportation Sector. In *Privacy Technologies and Policy SE - 10*, volume 8319, pages 157–176. 2014.

[127] Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.

[128] Alan Krueger, Alexandre Mas, and Xiaotong Niu. The evolution of rotation group bias: Will the real unemployment rate please stand up? Technical report, National Bureau of Economic Research, 2014.

[129] Alan B. Krueger, Andreas Mueller, Steven J. Davis, and Aysegul Sahin. Job search, emotional well-being, and job finding in a period of mass unemployment: Evidence from high frequency longitudinal data [with comments and discussion]. *Brookings Papers on Economic Activity*, pages pp. 1–81, 2011.

[130] P R Krugman. *The Self Organizing Economy*. Blackwell Publishers, 1996.

[131] John Krumm, Eric Horvitz, Paul Dourish, and Adrian Friday. Predestination: Inferring Destinations from Partial Trajectories. *UbiComp 2006: Ubiquitous Computing*, 4206:243–260, 2006.

[132] Victor Lavrenko, Matt Schmill, Dawn Lawrie, Paul Ogilvie, David Jensen, and James Allan. Mining of concurrent text and time series. In *KDD-2000 Workshop on Text Mining*, pages 37–44. Citeseer, 2000.

[133] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. Big data. The parable of Google Flu: traps in big data analysis. *Science (New York, N.Y.)*, 343(6176):1203–5, March 2014.

[134] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Computational Social Science. *Science*, 323(5915):721–723, 2009.

[135] Kyunghan Lee Kyunghan Lee, Seongik Hong Seongik Hong, Seong Joon Kim Seong Joon Kim, Injong Rhee Injong Rhee, and Song Chong Song Chong. SLAW: A New Mobility Model for Human Walks. *IEEE INFOCOM 2009*, 2009.

[136] David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628, 2005.

[137] Tong Liu, Paramvir Bahl, and Imrich Chlamtac. Mobility modeling, location tracking, and trajectory prediction in wireless ATM networks. *IEEE Journal on Selected Areas in Communications*, 16:922–935, 1998.

[138] HP Lo, N Zhang, and William HK Lam. Estimation of an origin-destination matrix with random link choice proportions: a statistical approach. *Transportation Research Part B: Methodological*, 30(4):309–324, 1996.

[139] Chung-Cheng Lu, Xuesong Zhou, and Kuilin Zhang. Dynamic origin–destination demand flow estimation under congested traffic conditions. *Transportation Research Part C: Emerging Technologies*, 34:16–37, 2013.

[140] Xin Lu, Linus Bengtsson, and Petter Holme. Predictability of population displacement after the 2010 Haiti earthquake. *Proceedings of the National Academy of Sciences of the United States of America*, 109(29):11576–81, July 2012.

[141] Xin Lu, Linus Bengtsson, and Petter Holme. Predictability of population displacement after the 2010 haiti earthquake. *Proceedings of the National Academy of Sciences*, 109(29):11576–11581, 2012.

[142] Ingrid Lunden. Austin's other event: A class action, mobile app privacy lawsuit filed against facebook, twitter, apple, 15 others, March 2012.

[143] Kees Maat, Bert van Wee, and Dominic Stead. Land use and travel behaviour: Expected effects from the perspective of utility theory and activity-based theories. *Environment and Planning B: Planning and Design*, 32:33–46, 2005.

[144] MJ Maher. Inferences on trip matrices from observations on link volumes: a bayesian statistical approach. *Transportation Research Part B: Methodological*, 17(6):435–447, 1983.

[145] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. Big data: The next frontier for innovation, competition, and productivity. Technical report, McKinsery & Company, May 2011.

[146] D Marcuss and Richard E Kane. Us national income and product statistics born of the great depression and world war ii. Technical report, Bureau of Economic Analysis, 2007.

[147] Mark Mazzetti and Michael S. Schmidt. Edward snowden, ex-c.i.a. worker, says he disclosed u.s. surveillance. *The New York Times*, June 2013.

[148] Aleecia M. McDonald and Lorrie Faith Cranor. Cost of reading privacy policies, the. *I/S: A Journal of Law and Policy for the Information Society*, 4:543, 2008.

[149] Matt McGee. Google has stopped street view photography in germany, April 2011.

[150] Michael G. McNally. The Four Step Model. *Center for Activity Systems Analysis*, November 2008.

[151] Sandro Meloni, Nicola Perra, Alex Arenas, Sergio Gómez, Yamir Moreno, and Alessandro Vespignani. Modeling human mobility responses to the large-scale spreading of infectious diseases. *Scientific reports*, 1:62, January 2011.

[152] Deepak K Merchant and George L Nemhauser. A model and an algorithm for the dynamic traffic assignment problems. *Transportation Science*, 12(3):183–199, 1978.

[153] Michelle N. Meyer. Everything you need to know about facebook's controversial emotion experiment, June 2014.

[154] Costas Milas and Philip Rothman. Out-of-sample forecasting of unemployment rates with pooled STVECM forecasts. *International Journal of Forecasting*, 24(1):101–121, January 2008.

[155] Claire Cain Miller. Revelations of n.s.a. spying cost u.s. tech companies. *The New York Times*, March 2014.

[156] Nicole Miller and Hugo Gaouette. Saudi arabia, RIM reportedly agree on BlackBerry use, u.s. says, August 2010.

[157] Kim Minkyong, David Kotz, and Kim Songkuk. Extracting a mobility model from real user traces. In *Proceedings - IEEE INFOCOM*, 2006.

[158] Mike Mintz. Class action lawsuit against app developers accuses them of stealing personal data, March 2012.

[159] Giovanna Miritello, Rubén Lara, Manuel Cebrian, and Esteban Moro. Limited communication capacity unveils strategies for human interaction. *Scientific reports*, 3, 2013.

[160] Alan L Montgomery, Victor Zarnowitz, Ruey S Tsay, and George C Tiao. Forecasting the us unemployment rate. *Journal of the American Statistical Association*, 93(442):478–493, 1998.

[161] M E J Newman and Juyong Park. Why social networks are different from other types of networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 68(3 Pt 2):036122, 2003.

[162] Mark EJ Newman. A measure of betweenness centrality based on random walks. *Social networks*, 27(1):39–54, 2005.

[163] Christos Nicolaides, Luis Cueto-Felgueroso, Marta C. González, and Ruben Juanes. A metric of influential spreading during contagion dynamics through the air transportation network. *PLoS ONE*, 7, 2012.

[164] Yu Nie, HM Zhang, and WW Recker. Inferring origin–destination trip matrices with a decoupled gls path flow estimator. *Transportation Research Part B: Methodological*, 39(6):497–518, 2005.

[165] Philip Oltermann. Germany opens inquiry into claims NSA tapped angela merkel's phone, June 2014.

[166] J-P Onnela, J Saramäki, J Hyvönen, G Szabó, D Lazer, K Kaski, J Kertész, and A-L Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences of the United States of America*, 104(18):7332–7336, 2007.

[167] Jukka-Pekka Onnela, Samuel Arbesman, Marta C González, Albert-László Barabási, and Nicholas A Christakis. Geographic Constraints on Social Network Groups. *PLoS ONE*, 6(4):7, 2011.

[168] Philip Oreopoulos, Marianne Page, and Ann Huff Stevens. The intergenerational effects of worker displacement. *Journal of Labor Economics*, 26(3):0–455, 2008.

[169] J de Ortúzar and Luis G Willumsen. *Modelling transport*. John Wiley & Sons, Chichester, England, 1994.

[170] Juan de Dios Ortúzar and Luis G. Willumsen. *Modelling Transport*. 2011.

[171] Wei Pan, Gourab Ghoshal, Coco Krumme, Manuel Cebrian, and Alex Pentland. Urban characteristics attributable to density-driven tie formation. *Nature communications*, 4:1961, 2013.

[172] Santi Phithakkitnukoon, Teerayut Horanont, Giusy Di Lorenzo, Ryosuke Shibasaki, and Carlo Ratti. Activity-aware map: Identifying human daily activity pattern using mobile phone data. *Human Behavior Understanding*, pages 14–25, 2010.

[173] Gyan Ranjan, Hui Zang, Zhi-Li Zhang, and Jean Bolot. Are call detail records biased for sampling human mobility? *ACM SIGMOBILE Mobile Computing and Communications Review*, 16(3):33–44, 2012.

[174] Carlo Ratti, Riccardo Maria Pulselli, Sarah Williams, and Dennis Frenchman. Mobile Landscapes: using location data from cell phones for urban analysis, 2006.

[175] J. Reades, F. Calabrese, and C. Ratti. Eigenplaces: analysing cities using the spaceÿ - ÿtime structure of the mobile phone network. *Environment and Planning B: Planning and Design*, 36(5):824–836, 2009.

[176] Jonathan Reades, Francesco Calabrese, and Carlo Ratti. Eigenplaces: analysing cities using the space-time structure of the mobile phone network. *Environment and Planning B: Planning and Design*, 36(5):824–836, 2009.

[177] Don Reisinger. Federal court OKs warrantless cell phone tracking by police, August 2012.

[178] Anthony J Richardson, Elizabeth S Ampt, and Arnim H Meyburg. *Survey methods for transport planning*. Eucalyptus Press Melbourne, 1995.

[179] Francesca Robin. The emerging market that could kill the iPhone, August 2012.

[180] Barbara Rossi. Do DSGE Models Forecast More Accurately Out-of-Sample than VAR Models? *Advances in Econometrics*, 32:27–79, 2013.

[181] Camille Roth, Soong Moon Kang, Michael Batty, and Marc Barthélemy. Structure of urban movements: Polycentric activity and entangled hierarchical flows. *PLoS ONE*, 6, 2011.

[182] Christopher J Ruhm. Are workers permanently scarred by job displacements? *The American Economic Review*, pages 319–324, 1991.

[183] Diego Rybski, Sergey V Buldyrev, Shlomo Havlin, Fredrik Liljeros, and Hernán A Makse. Scaling laws of human interaction activity. *Proceedings of the National Academy of Sciences*, 106(31):12640–12645, 2009.

[184] Adam Sadilek and John Krumm. Far Out: Predicting Long-Term Human Mobility. *AAAI*, pages 814–820, 2012.

[185] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[186] Samitha Samaranayake, Sebastien Blandin, and Alexandre Bayen. Learning the dependency structure of highway networks for traffic forecast. In *Proceedings of the IEEE Conference on Decision and Control*, pages 5983–5988, 2011.

[187] Paolo Santi, Giovanni Resta, Michael Szell, Stanislav Sobolevsky, Steven Strogatz, and Carlo Ratti. Taxi pooling in New York City: a network-based approach to social sharing problems. page 12, October 2013.

[188] Salvatore Scellato, Mirco Musolesi, Cecilia Mascolo, Vito Latora, and Andrew T. Campbell. NextPlace: A spatio-temporal prediction framework for pervasive systems. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6696 LNCS, pages 152–169, 2011.

[189] Michael S. Schmidt. Racy photos were often shared at n.s.a., snowden says. *The New York Times*, July 2014.

[190] Christian M Schneider, Vitaly Belik, Thomas Couronné, Zbigniew Smoreda, and Marta C González. Unravelling daily human mobility motifs. *Journal of the Royal Society, Interface / the Royal Society*, 10(84):20130246, 2013.

[191] Christian M Schneider, Vitaly Belik, Thomas Couronné, Zbigniew Smoreda, and Marta C González. Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, 10(84):20130246, 2013.

[192] Scott Schuh. An evaluation of recent macroeconomic forecast errors. *New England Economic Review*, pages 35–56, 2001.

[193] Robert P Schumaker and Hsinchun Chen. Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems (TOIS)*, 27(2):12, 2009.

[194] K Schwab, A Marcus, JO Oyola, W Hoffman, and M Luzi. *Personal data: The emergence of a new asset class*. The World Economic Forum, 2011.

[195] Mathew J. Schwartz. Sony data breach cleanup to cost $171 million, May 2011.

[196] Larry Seltzer. Your FaceBook data is FaceBooks's too, and forever, February 2009.

[197] Somini Sengupta. Europe moves to protect online privacy. *The New York Times*, February 2012.

[198] Andres Sevtsuk and Carlo Ratti. Does urban mobility have a daily routine? learning from the aggregate data of mobile networks. *Journal of Urban Technology*, 17(1):41–60, 2010.

[199] Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PloS one*, 6(5):e19467, 2011.

[200] Filippo Simini, Marta C González, Amos Maritan, and Albert-László Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):8–12, 2012.

[201] Michael E Smith. Design of small-sample home-interview travel surveys. *Transportation Research Record*, 701:29–35, 1979.

[202] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, September 2010.

[203] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, September 2010.

[204] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, 2010.

[205] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[206] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[207] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[208] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of Predictability in Human Mobility. *Science*, 327(5968):1018–1021, February 2010.

[209] V'ıctor Soto and Enrique F. Mart'ınez. Automated land use identification using cell-phone records. In *Proceedings of the 3rd ACM international workshop on MobiArch*, HotPlanet '11, pages 17–22, New York, NY, USA, 2011. ACM.

[210] Heinz Spiess. A maximum likelihood model for estimating origin-destination matrices. *Transportation Research Part B: Methodological*, 21(5):395–412, 1987.

[211] Heinz Spiess. Technical NoteÑConical Volume-Delay Functions. *Transportation Science*, 24(2):153–158, May 1990.

[212] Heinz Spiess. Technical noteâĂŤconical volume-delay functions. *Transportation Science*, 24(2):153–158, 1990.

[213] James H Stock and Mark W Watson. A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series. Technical report, National Bureau of Economic Research, June 1998.

[214] Peter R Stopher and Stephen P Greaves. Household travel surveys: Where are we going? *Transportation Research Part A: Policy and Practice*, 41(5):367–381, 2007.

[215] Tanya Suhoy. *Query indices and a 2008 downturn: Israeli data.* Research Department, Bank of Israel, 2009.

[216] Daniel Sullivan and Till Von Wachter. Job displacement and mortality: An analysis using administrative data. *The Quarterly Journal of Economics*, 124(3):1265–1306, 2009.

[217] L. Sun, K. W. Axhausen, D.-H. Lee, and X. Huang. Understanding metropolitan patterns of daily encounters. *Proceedings of the National Academy of Sciences*, 110(34):13774–13779, August 2013.

[218] Arvind Thiagarajan, Lenin Ravindranath, Katrina LaCurts, Samuel Madden, Hari Balakrishnan, Sivan Toledo, and Jakob Eriksson. VTrack: accurate, energy-aware road traffic delay estimation using mobile phones. In *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems - SenSys '09*, pages 85–98, 2009.

[219] Richard Tiller and Michael Welch. Predicting the National unemployment rate that the "old" CPS would have produced. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1994.

[220] Jameson L Toole, Meeyoung Cha, and Marta C González. Modeling the adoption of innovations in the presence of geographic and media influences. *PLoS ONE*, 7(1):e29528, 2012.

[221] J Ugander, L Backstrom, C Marlow, and J Kleinberg. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, 2012(16):1–5, 2012.

[222] Pauline van den Berg, Theo A. Arentze, and Harry J. P. Timmermans. Size and Composition of Ego-Centered Social Networks and Their Effect on Geographic Distance and Contact Frequency, 2010.

[223] Henk J Van Zuylen and Luis G Willumsen. The most likely trip matrix estimated from traffic counts. *Transportation Research Part B: Methodological*, 14(3):281–293, 1980.

[224] Alessandro Vespignani. Predicting the behavior of techno-social systems. *Science*, 325(5939):425–8, 2009.

[225] Alessandro Vespignani. Modelling dynamical processes in complex socio-technical systems, 2011.

[226] Katy Waldman. Facebook's unethical experiment. *Slate*, June 2014.

[227] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-László Barabási. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*, page 1100. ACM Press, 2011.

[228] Jingyuan Wang, Yu Mao, Jing Li, Chao Li, Zhang Xiong, and Wen-Xu Wang. Predictability of road traffic and congestion in urban areas. July 2014.

[229] Pu Wang, Timothy Hunter, Alexandre M Bayen, Katja Schechtner, and Marta C González. Understanding road usage patterns in urban areas. *Scientific reports*, 2:1001, January 2012.

[230] Pu Wang, Timothy Hunter, Alexandre M Bayen, Katja Schechtner, and Marta C González. Understanding road usage patterns in urban areas. *Scientific reports*, 2, 2012.

[231] Pu Wang, Like Liu, Xiamiao Li, Guanliang Li, and Marta C González. Empirical study of long-range connections in a road network offers new ingredient for navigation optimization models. *New Journal of Physics*, 16(1):013012, January 2014.

[232] John Glen Wardrop. Road paper. some theoretical aspects of road traffic research. In *ICE Proceedings: Engineering Divisions*, volume 1, pages 325–362. Thomas Telford, 1952.

[233] Duncan J. Watts. Lessons learned from the facebook study, July 2014.

[234] Duncan J Watts, Peter Sheridan Dodds, and M E J Newman. Identity and search in social networks. *Science (New York, N.Y.)*, 296(5571):1302–1305, 2002.

[235] Duncan J Watts, Roby Muhamad, Daniel C Medina, and Peter S Dodds. Multiscale, resurgent epidemics in a hierarchical metapopulation model. *Proceedings of the National Academy of Sciences of the United States of America*, 102(32):11157–11162, 2005.

[236] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of Ôsmall-worldÕ networks. *Nature*, 393(6684):440–442, 1998.

[237] Amy Wesolowski, Nathan Eagle, Abdisalan M Noor, Robert W Snow, and Caroline O Buckee. The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of the Royal Society, Interface / the Royal Society*, 10(81):20120986, April 2013.

[238] Amy Wesolowski, Nathan Eagle, Andrew J Tatem, David L Smith, Abdisalan M Noor, Robert W Snow, and Caroline O Buckee. Quantifying the impact of human mobility on malaria. *Science (New York, N.Y.)*, 338(6104):267–70, October 2012.

[239] Amy Wesolowski, Nathan Eagle, Andrew J Tatem, David L Smith, Abdisalan M Noor, Robert W Snow, and Caroline O Buckee. Quantifying the impact of human mobility on malaria. *Science*, 338(6104):267–270, 2012.

[240] William C. Wheaton. Land use and density in cities with congestion. *Journal of Urban Economics*, 43:258–272, 1998.

[241] Paul Wicks, Timothy E. Vaughan, Michael P. Massagli, and James Heywood. Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. *Nature Biotechnology*, 29(5):411–414, May 2011.

[242] Yingxiang Yang, Carlos Herrera, Nathan Eagle, and Marta C González. Limits of predictability in commuting flows in the absence of data for calibration. *Scientific reports*, 4:5662, January 2014.

[243] Jeffrey J. Yankow. Why do cities pay more? An empirical examination of some competing theories of the urban wage premium. *Journal of Urban Economics*, 60:139–161, 2006.

[244] Jing Yuan, Yu Zheng, and Xie Xing. Discovering regions of different functions in a city using human mobility and pois. *KDD*, 2012.

[245] João Zamite, Fabrício A B Silva, Francisco Couto, and Mário J Silva. MED-Collector: Multisource epidemic data collector. In *Transactions on large-scale data-and knowledge-centered systems IV*, pages 40–72. Springer, 2011.

[246] Xianyuan Zhan, Samiul Hasan, Satish V Ukkusuri, and Camille Kamga. Urban link travel time estimation using large-scale taxi data with partial information. *Transportation Research Part C: Emerging Technologies*, 33:37–49, 2013.

[247] Yu Zheng and Xing Xie. Learning travel recommendations from user-generated gps traces. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(1):2, 2011.