Inferring Asymmetry of Inhabitant Flow using Call Detail Records

senseable city lab:.::

Inferring Asymmetry of Inhabitant Flow using Call Detail Records

Santi Phithakkitnukoon and Carlo Ratti
SENSEable City Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
Email: {santi, ratti}@mit.edu

Abstract— **In this research, we carry out a study of the inhabitant flow using a large mobile phone data with location estimates from subscribers in Suffolk county, Massachusetts, USA that reveals the asymmetry in the flows, which reflects the way that people travel daily. People occasionally travel in a non-symmetrical way. For instance, they would take one route traveling from home to a destination and a different route while returning home. By analyzing the flow over the space, the results show that there exists asymmetrical flows, which account for 33% of all inhabitant flows. In addition, high asymmetrical flows are observed in trips between low and high congested areas e.g. urban and suburban areas, as well as trips made to and from low populated areas e.g. commercial areas.**

Index Terms— **Inhabitant flow, mobile phone data mining, urban computing**

I. Introduction

Spatial interaction processes play an important role in planning and design of urban area and public transportation. Spatial interaction has been studied in an international scale [36] as well as national scale [50] [29] [30] [31]. The interaction is characterized by the "flow," which has referred to the number of passengers in a transportation system such as trains, buses, and airways [26], number of phone messages [26], number of travelers in road networks [25] [5], and density of the calls in phone communication networks [27].

Though these studies have provided better understanding of inter-nation/city flows, a finer analysis at a micro level is yet still difficult due to the lack of data. Recently, with the large datasets made available by academic institutions ([37] [41] [13]) and industries [3] [45], researchers are given the unique opportunity to analyze human mobility at a finer gained level [4] [28] [9] [16] [46]. With the rapidly increase of mobile phone users, mobile phones become the sensors of the city – sensing social networks [38] [39] [16] [13], social events [8], and activities [40].

While inter-city/urban system spatial interaction has been widely studied ([36] [50] [29] [30] [31]), to date research studies have not yet analyzed spatial interaction within the city at kilometer-square scale. In this study, the use of the graphical information from the large mobile phone data allows us to investigate the human mobility within the city from which reveals the symmetry in inhabitant flow and asymmetry in traveling pattern. The key findings of this research are:

(1) Asymmetrical traveling pattern: With our large mobility data, we are able to capture a symmetrical inhabitant flow that describes the core element of daily mobility pattern in the way that people begin and end their trips at the same locations e.g. homes. We find that people travel in an asymmetrical way i.e. for each trip between home and some destination, people don't always take the completely same routes traveling to and returning from the destinations – route choice changes from one end to the other. We have also observed that these asymmetrical flows account for 33% of entire inhabitant flows – meaning that one third of the flows in Suffolk county, MA (county of study) are imbalanced.

(2) Intensity of asymmetrical flows: The flows with high degree of asymmetry are observed from the trips between high and low congested areas e.g. urban and suburban areas, as well as trips made to and from low populated areas e.g. commercial areas.

The rest of the paper is organized as follows. In Section II, we briefly review some literature in spatial interaction and route choice behavior that are relevant to our work. We then describe our dataset and methodology used in this study of inhabitant flow in Section III, including within-cell flow, pairwise flow, and correlation analysis. Applications of our study in O-D matrix estimation are discussed in Section IV. Section V concludes the paper with summary and an outlook on future work.

II. Related Work

A. Spatial Interaction

To date, the studies of inhabitant flows have been done based on census and survey data at municipal level. Researchers have focused on the flows between home and work locations ([36] [50]) as well as single-day trips ([29] [30] [31]) across different urban areas.

Nielsen and Hovgesen [36] construct a origin-destination (OD) matrix using home and work location information from 1991 and 2001 census from which analysis of interaction patterns within England and Wales is performed. They divide the map into 5x5km^2 grids and define the flow as the number of commuters passing through – assuming that their paths are straight lines between their home and work locations. Their study is focused on geographical variations and a map of commuter flows within England and Wales.

Eck and Snellen [50] analyze the home-to-work commuting patterns within the Randstad Holland (a conurbation in the Netherlands). They measure the strength and symmetry of home-to-work flows based on data from the Dutch Travel Survey (OVG) and the Questionnaire Labour Force (EBB) from the periods 1990-1994 and 2001-2003. The data contains the information about home and work locations. The strength is measured as the number of commuters and asymmetry score is computed as the relative difference between flows. Their study aims to elaborate on the concept of city network within Randstad by analyzing asymmetry of commuting patterns between municipalities.

Limtanakool et al. [29] [30] [31] develop theoretical framework for spatial interaction (strength, symmetry, and structure), which are used to examine the change in the configuration of the urban systems on the basis of commute and leisure flows between 23 daily urban systems. Their results are based on the data obtained from 1992-99 and 2002 Netherlands National Travel Surveys (NTSs) that include information on the purpose, self-reported distance and time, mode, and the geographical location of origin and destination (measured in municipal level) of all trips for a single day.

In contrast to aforementioned works, our study is concerned with inhabitant flow in space produced by all types of trip (not only home-to-work commuting direction as studied in [50]) over extensive period of time (rather than single-day trips analyzed in [29] [30] [31]) with finer grained human mobility data. Our focus is on the flows throughout the urban space and symmetry property of the flows within the space that reflects the characteristics of inhabitant daily traveling patterns.

B. Route Choice Behavior

As our findings are concerned with the conservativeness of the traveling patterns that are caused by most likely the commuter's route choices, so we briefly review related literature in route choice behavior.

Hill [20] analyzes strategies used by pedestrians in selecting and describing routes and finds that route selection strategies are largely subconscious. Moreover, *directness* is the most common reason for choosing a particular route. The directness of the route does not only concern the length of the route, but also its complexity (in terms of changes in direction). Senevarante and Morall [44] and Guy [17] describe that pedestrians frequently choose the shortest route while not aware that minimizing distance is their primary strategy in choosing route. The factors such as the level of congestion, safety or visual attractions are only secondary. Bovy and Stern [7] indicate that together with distance, pleasantness is an important route attribute that produce a high correlation with the route preferences. Other attributes that influence route choice behavior include habit, number of crossings, pollution and noise levels, safety and shelter from poor weather conditions, and stimulation of the environment. These attributes influence to what extent has to do with the trip purpose, e.g., scenery plays a very important role in recreational trips but not in work-related walking trips. Helbing [19] has shown that there are some indications that pedestrians somehow optimize the order in which they perform their activities and that *directness* plays an important role.

Modeling pedestrian travel behavior has motivated a number of researchers. There are indeed several sources of uncertainty in a route choice context such as limited knowledge of the route network, uncertain perceived travel time, and unavailability of pedestrian's individual characteristics and preferences. The most widely used approach is the *utility model framework* where pedestrians or travelers are assumed to maximize utility (probability that a particular alternative route is chosen based on other attributes and a random term capturing uncertainty). Dial [11] was one of the first researchers who implement this model.

Lovas [33] uses *Markov model* to describe movement of pedestrian from one node of the network to another. *Queueing model* is then used to estimate the waiting times when pedestrian traffic demand is larger than the door capacity, and describe pedestrian evacuation behavior from buildings. Gipps [14] and Hamacher and Tjandra [18] apply basic *discrete choice modeling* to describe pedestrian route choice in walking facility with a finite number of routes. Verlander [51] estimates discrete choice models using household-based diary data. Teklenburg et al. [47] use a *Space Syntax model* calibrated utilizing pedestrian flow data.

Assuming a finite set of route alternatives may not be feasible in real life scenario where pedestrians normally choose the route from a countless number of possible routes. Hughes [24] accounts for this aspect and applies *potential function* to describe the optimal walking direction to the destination (in terms of travel time) as a function of the current location of the pedestrian. Hoogendoorn and Bovy [23] take a similar approach and propose a theoretical framework that also includes some general route attributes that are not considered by Hughes [24] such as walking distance, stimulation of the environment, and uncertainty in the traffic conditions expected by the pedestrians. Rich et al. [42] examine the possibility of using GPS data to provide information on alternative routes and suggest that route alternatives can be formed as a set of *sub-paths* defined as unions of sub-paths of GPS-logged routes. This way, sub-paths may be sampled and joined to form new alternative routes.

Pedestrian traffic and vehicular traffic are different in traffic operations and travel behavior. Nevertheless, the route choice behaviors are quite similar. Hence route choice models assume people (drivers) adhere to an underlying *utility function* when choosing a route. The internal process weighs up the costs of various aspects of the upcoming journeys and selects the journey that minimizes cost and maximizes utility. Duckham and Kulik [12] introduce the idea of “simplest” paths, which is the minimum angular path that might be useful for navigation system to direct drivers to destination more simply. Modeling movement of pedestrians and drivers has also motivated the filed of *Space syntax* [21]. Similar to the most network analysis, space syntax turns a network map into a graph (nodes representing lines while junctions are represented by links) and graph analysis is then performed to determine the movement. Therefore, space syntax is quite suitable for a route choice problem as turns increment the cost of a trip whereas straight-line segments do not. Turner [48] introduce *angular segment analysis* to space syntax in which the junctions and turns break the lines into segments. The cost of transfer from one segment to another is treated as the angle from one segment to another. Hochmair and Frank [22], Conroy and Dalton [10], and Turnerand and Dalton [49] have showed results that support the *shortest angular path hypothesis*, which states that people try to direct themselves according to the (perceived) current minimum angle to the destination.

People take the shortest path as their primary route choice strategy. The shortest path as perceived, however, may not be the actual shortest path calculated in physical distance. A route that seems shorter or quicker or straighter from one end may not be so perceived from the other end, thus inducing a change of route [15]. Rather than analyzing the individual route choice behavior, in this study, we analyze route choice in forms of the collective mobility flux throughout the space and determine the underlying behavior as a whole.

III. Inhabitant Flow

In this study, we use anonymous mobile phone data collected in October 2009 by Airsage [3] of one million users in the state of Massachusetts, USA, which account for approximately 20% of population, equally spread over space. This includes 130 million anonymous location estimations in (latitude,longitude)-coordinates, which were recorded when the users were engaged in communication via the cellular network. Specifically, the locations are estimated at the beginning and the end of each voice call placed or received, when a short message is sent or received, and while internet is connected. Note that these location estimations have an average uncertainty of 320 meters and median of 220 meters as reported by Airsage [3] based on internal and independent tests.

Due to the limitation in the location estimates that are only available when a communication is engaged and our interest in fine grained mobility patterns, we have selected top 3,000 users who is the most active users and subscribers of the phone service providers in Suffolk county (whose seat is Boston – the state capital, and the largest city). To ensure our subject’s mobility, we further select only the subjects whose home and work locations are different. Home and work locations are estimated as the locations in which the subjects occupy most frequently during the night and day hours, respectively. As the result, we obtain 838 mobile phone users as the subjects of this study, which account for more than two million location traces. The data is then resampled at a constant rate of 10 minutes (in the other words, the location estimates are available every 10 minutes). The location traces of these mobile phone users are plotted in Fig. 1.

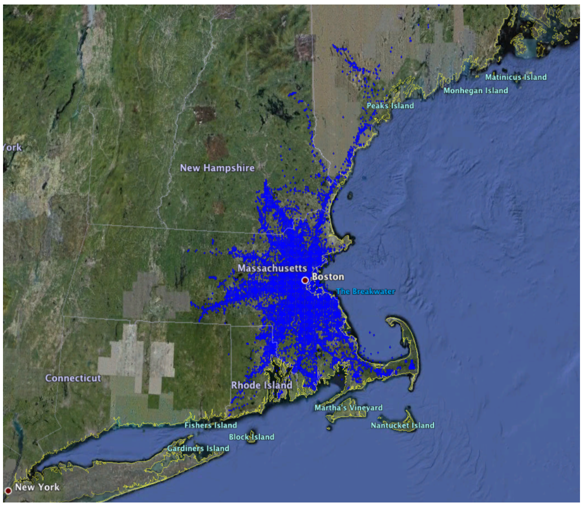


Figure 1. Location traces collected in in October 2009 of mobile phone users who are subscribers of the phone service providers in Suffolk county whose seat is Boston, the Massachusetts state capital and the largest city.

To compensate for the uncertainty of location estimates, we model the map with the virtual 500-meter square grid cells. The traces are then characterized by these virtual cells. In addition to the uncertainty in location estimations, there are some outliers such as locations reported in the ocean, irregular remote places, etc. A common characteristic of these outliers is that they are single isolated points. So, to remove these outliers, we filter out all single flows between the cells.

With our preprocessed data, in the next subsections we will describe our analyses of the inhabitant flow characterized by the mobility patterns of mobile phone users in Suffolk. The analyses will be carried out from two aspects: within-cell and pairwise flows.

A. *Within-cell Flow*

In this subsection, we will investigate on the flow within the cell. The *flow* here is defined as a number of people who are moving. If the direction is outward (i.e. going out of the cell), then it is the *outgoing flow*. Conversely, if the direction is inward, then the flow is referred as the *incoming flow*. We are interested in the flow within each of 500-meter square cells in the map. In particular, our focus is in the amount, direction, and symmetry of the flow of each cell as well as its spatial characteristics associated with it.

With our 500-meter square grid cells, we construct the *transition matrix* M that contains elements $m(i,j)$ where $i,j \in \{1,2,3,...,N\}$ representing the flow between cells

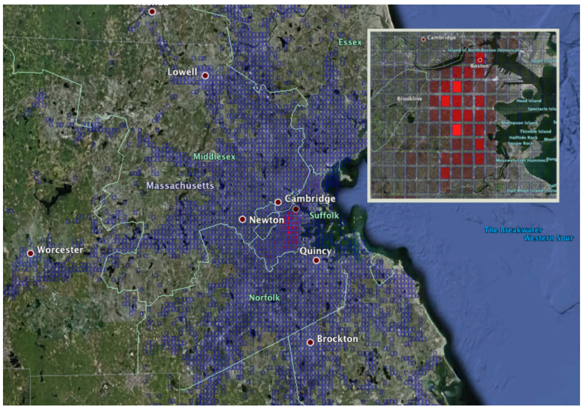


Figure 2. Higher flows can be observed near Boston area where the value of the flow is indicated by the reddish shad. Zoom-in is the Boston area with the high flows.

as follow:

$$M = \begin{bmatrix} m(1,1) & m(1,2) & \cdots & m(1,N) \\ m(2,1) & m(2,2) & \cdots & m(2,N) \\ \vdots & \vdots & \ddots & \vdots \\ m(N,1) & m(N,2) & \cdots & m(N,N) \end{bmatrix} \quad (1)$$

where $m(i,j)$ is the flow from cell i to cell j and N is the total number of visited cells. Based on our data, the users visited 2,706 different cells (i.e. $N = 2,706$).

For each cell c where $c \in \{1,2,3,...,N\}$, the incoming flow ($f_{in}(i)$) of cell i is defined as the total number of people entering cell i from any cell in the map:

$$f_{in}(i) = \sum_{k=1,k\neq i}^{N} m(k,i), \quad (2)$$

and the outgoing flow ($f_{out}(i)$) of cell i is defined similarly as the total number of people going from cell i to any cell in the map:

$$f_{out}(i) = \sum_{k=1,k\neq i}^{N} m(i,k). \quad (3)$$

We compute the incoming and outgoing flows for each cell and find that *the flow within each cell is symmetrical*, which means that the incoming flow is equal to outgoing flow for every cell, i.e. $\forall i : f_{in}(i) = f_{out}(i)$. The high-flow cells appear to be near Boston area, which is the most crowded area (see Fig. 2). The distribution of the flows is shown in Fig. 3 where its value is as high as 11,380 and as low as 2 with mode, mean, and median of 2.00, 294.20, and 12.00 respectively.

The discovered symmetrical flow essentially means that people go to places and eventually return to where they are originally. In the other words, people begin and end their trips at the same locations, e.g. homes. This result also tells us that our data is fine-grained, from which the routine behavioral patterns such as commuting flows with respect to home locations can be extracted precisely. As our subjects are subscribers of the phone service providers in Suffolk, so it is very likely that they reside in Boston and surrounded areas, hence the result is sensible.

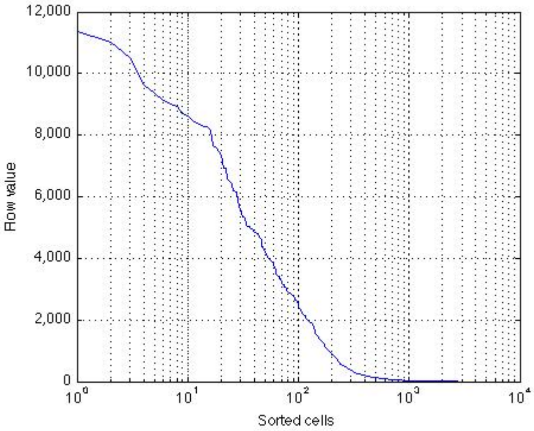


Figure 3. Distribution of the flows.

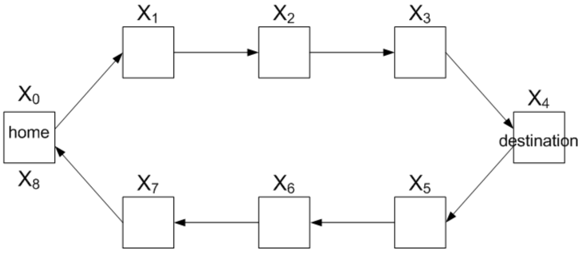


Figure 4. An example of a trip – a cyclic graph.

B. Pairwise Flow

In this subsection, we explore further about the inhabitant flow by considering the flow between cells ($m(i,j)$). Since the result in the previous section shows that the within-cell flow is symmetrical, we thus examine the pairwise flows to determine whether or not the symmetry condition also holds.

Before finding out about the symmetry of pairwise flow, we would like to introduce some definitions, which will be later used. Suppose a user's *trace* is defined as $X = \{x_1 \rightarrow x_2 \rightarrow \ldots \rightarrow x_k \rightarrow \ldots\}$ where x_k = [latitude, longitude, timestamp]. According to the result from the previous section that people travel symmetrically with respect to home locations, a *trip* can then be defined as a set of x's: $T_s = \{x_k \rightarrow x_{k+1} \rightarrow x_{k+2} \rightarrow \ldots \rightarrow x_{k+n}\}$ where $(lat,long)_k = (lat,long)_{k+n}$ i.e. home. Hence each trip is a cyclic graph – a round-trip graph that begins and ends at home location as shown in Fig. 4. Each subject makes several trips throughout the course of observation period i.e. $s = 1,2,3,,....$

According to Fig. 4, the *forward trip* is defined as a set of x_k going from home to destination: $T_s^{(forward)} = \{x_0 \rightarrow x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow x_4\}$. Likewise, the *backward trip* is defined as a set of x_k coming back home from destination: $T_s^{(backward)} = \{x_4 \rightarrow x_5 \rightarrow x_6 \rightarrow x_7 \rightarrow x_8\}$.

A trip can classified into two types: symmetrical and asymmetrical. A *symmetrical trip* is a trip where $T^{(forward)}$ and $T^{(backward)}$ are completely overlapped in (lat,long)-coordinates as shown in Fig. 5. In the other

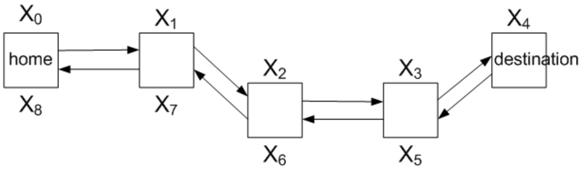


Figure 5. An example of a symmetrical trip where $T^{(forward)}$ and $T^{(backward)}$ are completely overlapped in (lat,long)-coordinates.

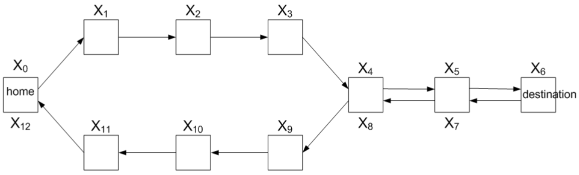


Figure 6. An example of an asymmetrical trip where $T^{(forward)}$ and $T^{(backward)}$ are not completely overlapped in (lat,long)-coordinates.

words, the traveler uses the same routes in forward and backward direction i.e. route choice does not change. On the other hand, the *asymmetrical trip* is defined as a trip where $T^{(forward)}$ and $T^{(backward)}$ are not completely overlapped in (lat,long)-coordinates as shown in Fig. 6. Hence the traveler's route choice changes from one end to the other.

Using the developed concepts of the symmetrical/asymmetrical trips and transition matrix, we form the following definitions:

Definition 1: If $\forall i, j : m(i, j) = m(j, i)$, then all trips are symmetrical.

Definition 2: If $\exists i, j : m(i, j) \neq m(j, i)$, then there exists asymmetrical trips.

Here, we are basically trying determine whether or not M is a symmetric matrix, i.e. $M = M^T$ or $\forall i, j : m(i, j) = m(j, i)$. It however turns out that some elements in M do not match across the diagonal. Hence M is not symmetric i.e. $\exists i, j : m(i, j) \neq m(j, i)$. Therefore, according to Def. 2, there exists asymmetrical trips, which means that the travelers did not always use the same routes traveling between home and other places in forward and backward directions. *Hence there is some degree of asymmetry associated with their traveling patterns.*

To quantify the asymmetry level of the flow, the *difference matrix D* is constructed. It contains the elements that denote the differences in the flows between cell i and j as follows:

$$D = \begin{bmatrix} d(1,1) & d(1,2) & \cdots & d(1,N) \\ - & d(2,2) & \cdots & d(2,N) \\ - & - & \ddots & \vdots \\ - & - & - & d(N,N) \end{bmatrix} \tag{4}$$

where $d(i, j) = |m(i, j) - m(j, i)|$. Since lower triangular matrix is just a mirror of the upper triangular matrix ($\forall i, j : d(i, j) = d(j, i)$), one side of the matrix is only required for constructing D.

The next step is to extract all non-zero elements in D, which exhibit the asymmetric flows. The total of 15,275 elements are found to have non-zero values in D. The distribution of these asymmetric flows is shown in Fig. 7 where the maximum and minimum values are 128 and 1 respectively while mean, mode, and median are 3.44, 2.00, and 2.00 respectively.

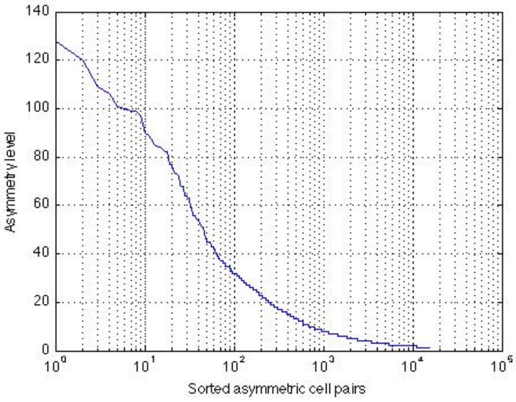


Figure 7. Distribution of the asymmetric flows.

The non-zero elements in D represent the absolute values of asymmetric flows. It tells us the density of the asymmetric flows but does not provide the value that is relative to the overall flows between cells. Thus, we define the *relative asymmetry level* of the flows between cell i and j as the elements of the *relative difference matrix* D_R as follows:

$$D_R = \begin{bmatrix} d_R(1,1) & d_R(1,2) & \cdots & d_R(1,N) \\ - & d_R(2,2) & \cdots & d_R(2,N) \\ - & - & \ddots & \vdots \\ - & - & - & d_R(N,N) \end{bmatrix} \tag{5}$$

where $d_R(i, j) = \frac{d(i,j)}{max(m(i,j),m(j,i))}$. Similar to D, since lower triangular matrix is just a mirror of the upper triangular matrix ($d_R(i, j) = d_R(j, i)$), one side of the matrix across the diagonal is only required. The value of $d_R(i, j)$ is within the range of $[0, 1]$ where 0 implies symmetric flow and 1 implies the maximum relative asymmetric flow, which also means the unidirectional flow.

The relative asymmetry level here represents the fraction of the asymmetry in the flows of a given cell pair. For example, if $m(a, b) = 100$ and $m(b, a) = 110$, then the relative asymmetry level would be $10/110 = 0.09$ or 9% with respect to the maximum flow, which implies that the flow is 9% away from being symmetrical. With this same absolute asymmetrical flow of 10 but $m(a, b) = 1,000$ and $m(b, a) = 1,010$ instead, the relative asymmetry level would be 1%. So the relative asymmetry value provides the insight on the quantity of asymmetrical flow being away from reaching symmetrical level.

The distribution of the relative asymmetrical flows is shown in Fig. 8 where the value is as high as 1.00 and as low as 0.002 while the values of mean, mode, and median are 0.60, 1.00, and 0.50 respectively.

As mentioned previously that this asymmetrical flow reflects the asymmetrical traveling behavior of the inhabitants. An interesting question is how much does

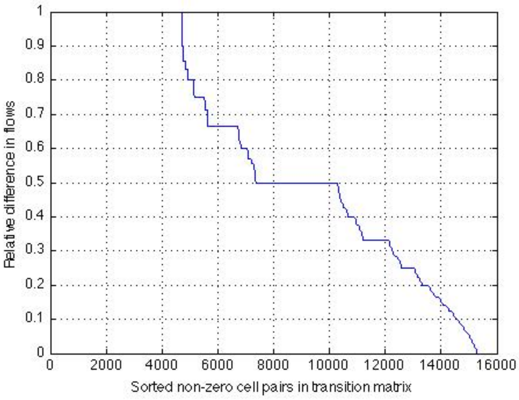


Figure 8. Distribution of the relative asymmetrical flows.

TABLE I.
CLASSIFIED GROUPS OF RELATIVE ASYMMETRIC FLOWS

Group	Low	Medium	High	Very high
Range	$0 < d_R \leq 0.3$	$0.3 < d_R \leq 0.6$	$0.6 < d_R < 1$	$d_R = 1$
Group Size	3,042	5,408	2,104	4,721

asymmetrical traveling behavior occur. To find out, we compute the percentage of asymmetry level ($Asym$) as

$$Asym = \frac{\sum\limits_{\forall i,j:i \neq j} |m_u(i,j) - m_u(j,i)|}{\sum\limits_{\forall i,j:i \neq j} max(m_u(i,j), m_u(j,i))}, \qquad (6)$$

where M_u is the upper triangular matrix of M with elements $m_u(i,j)$.

As the result, asymmetric flows account for 33.01% of the entire flows. Roughly speaking, *people change their route choice one third of their daily trips*. For instance, for any given 10 days, a person would commute from home to work with the exact same route back and forth for seven days and would take alternative routes either forward or backward direction for other three days.

To give the sense of geographical locations of these asymmetric flows, the relative asymmetric flows between 15,275 cell pairs are shown in Fig. 9. The color used to differentiate four different groups of the flows based on the relative asymmetry level. The group size (number of flows) and value range of each classified group is given in Table III-B. White, blue, yellow, and red represent low, medium, high, and very high flows respectively. Each group also is plotted separately in Fig. 10.

The low relative asymmetrical flows are clustered closely around the central Boston area, which is a crowded area with high numbers of inhabitant flows (as seen in previous section). Since the central Boston area has high inhabitant flows, it means that the area is a common traveling space that is very likely to comprise the routine traveling paths of many people. The route choice in routine trips is clearly less random than the non-routine trips. It however does not necessary mean that there are no asymmetrical traveling patterns in the routine trips. In fact, it appears that there are asymmetrical traveling patterns within the area but at a low degree compared to traveling patterns further away from the high inhabitant flow area.

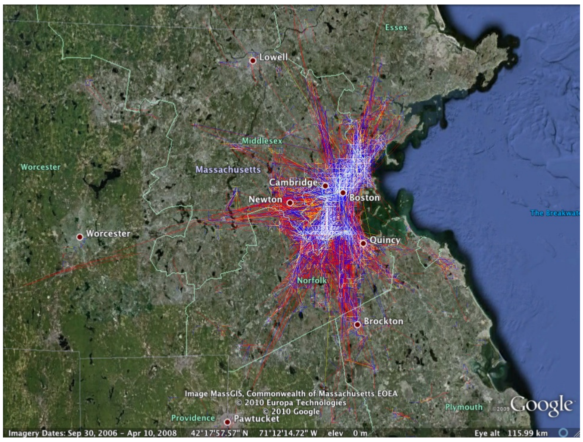


Figure 9. Geographical location of the relative asymmetric flows on the map. White, blue, yellow, and red represent low, medium, high, and very high flows respectively.

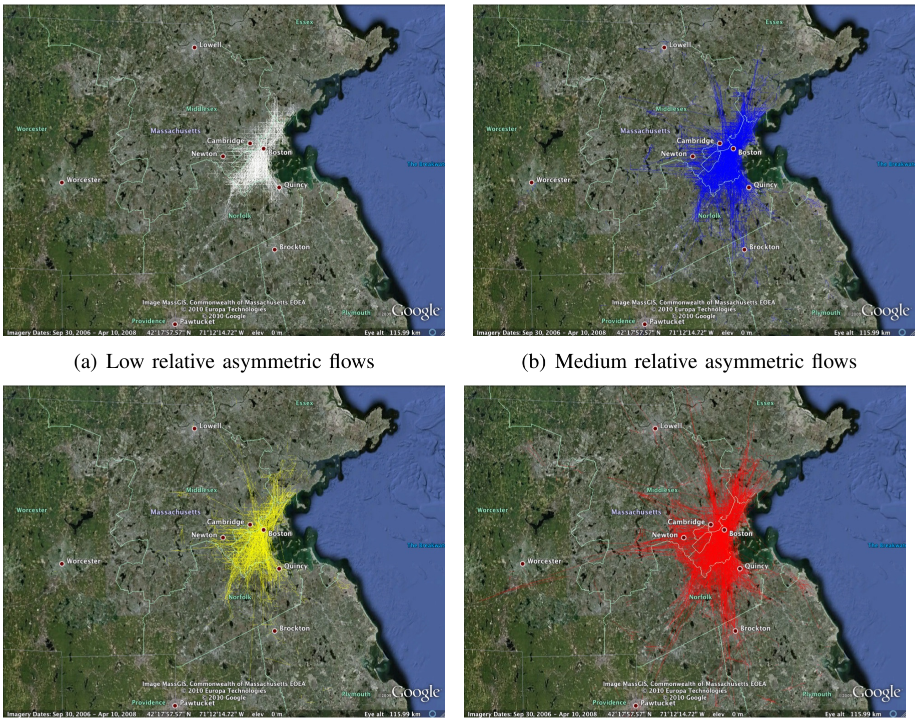
The medium and high relative asymmetrical flows are quite similar in the geographical area. The medium relative asymmetric flow group has the highest amount of flows among other groups. So it is fair to say that most asymmetrical traveling patterns that are not unidirectional trips are about 30–60% away from being symmetrical.

The very high relative asymmetric flows are unidirectional trips. They appear in a more spread area than other groups. With our fair assumption that more routine trips occur near central Boston area while less routine trips are further away, we can infer that both routine and non-routine trips consist of unidirectional flows. As an example, for routine trips such as home to work, some people consistently travel in an asymmetrical way by taking one path form home to work and another completely different path back home. This scenario creates the unidirectional flows in routine trips. On the other hand, when people travel to non-familiar locations (out of routine) that are typically far away from their normal living area, they tend to either explore the new places along the way such as restaurants and tourist spots, or are stumbling finding the way. Hence these scenarios would create unidirectional flows in non-routine trips.

The result is sensible and consistent with other researchers' findings in inhabitant's route choice behavior [43] [15] [49] [23] [17]. We believe that this study adds to their findings by focusing on the flow of inhabitants over extensive period of time.

C. Correlations

While the inhabitant flow of each cell reflects mobility flux, the relative asymmetry level describes the degree of the existence of asymmetrical traveling patterns of each cell pair. We therefore investigate further, in this section, on correlation between the within-cell flow and asymmetry level. In addition, we explore the relationship between the relative asymmetry level and population density as well as landuse categories.



(a) Low relative asymmetric flows

(b) Medium relative asymmetric flows

(c) High relative asymmetric flows

(d) Very high relative asymmetric flows

Figure 10. Geographical location of the relative asymmetric flows of each group.

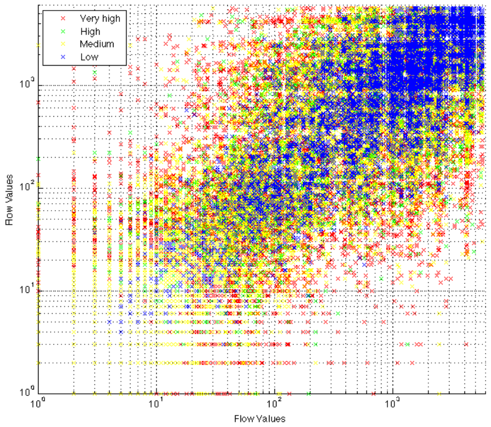


Figure 11. Relative asymmetry level associated with the flows between cells at different flow values.

In Fig. 11, a scatter plot of relative asymmetry levels (differentiated by color) associated with cell pairs of different flow values (described in Table III-B) is shown. The low asymmetry levels appear to be clustered around the flows between high flow cells while higher asymmetry levels are more associated with the flows of lower flow cells.

To be more precise in our observation from Fig. 11, we divide cells into three groups according to their flow levels, namely Low, Medium, and High. Table II shows flow range and size of each group. The average relative asymmetry level associated with the flows between these groups is plotted in Fig. 12 where, for example, L-M represents Low and Medium-flow cell pairs. It can be observed that the traveling flows between Low and High flow cells exhibit the highest average relative asymmetry level at about 0.80. This implies that *the asymmetrical traveling patterns tend to occur the most when the trips are between high and low congested areas*. It can also mean the trips between the urban and suburban areas. Details of each group is given in Table III.

TABLE II.
CLASSIFIED CELL GROUPS BASED ON FLOW VALUE

Group	Range	Group size	
		As a departing cell	As an entering cell
Low	$1 \leq f < 10^2$	4,394	5,346
Medium	$10^2 \leq f < 10^3$	5,225	5,597
High	$10^3 < f$	5,656	4,332

TABLE III.
AVERAGE RELATIVE ASYMMETRY LEVEL ALONG WITH THE NUMBER OF CELL PAIRS OF EACH GROUP WHERE L=LOW, M=MEDIUM, H=HIGH

Group of Cell Pairs	Group Size	Avg. Relative Asymmetry Level
L-L	3,551	0.72
L-M	2,149	0.68
L-H	489	0.80
M-M	2,605	0.54
M-H	3,463	0.59
H-H	3,018	0.45

To ensure that the result in Fig. 12 is not impacted by the errors or outliers in the data, we filter out traces that are longer than t km (threshold). We vary the threshold

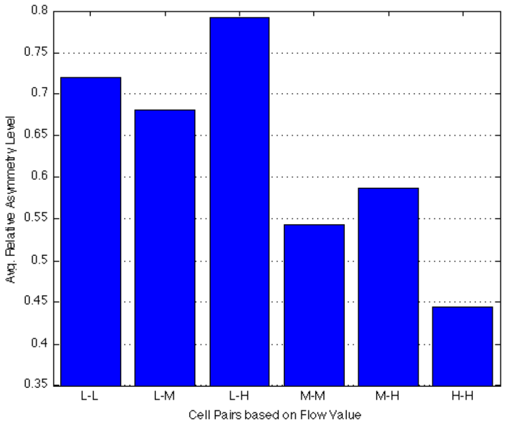


Figure 12. Average relative asymmetry level for different groups of cell pairs based on flow level where L=Low, M=Medium, and H=High.

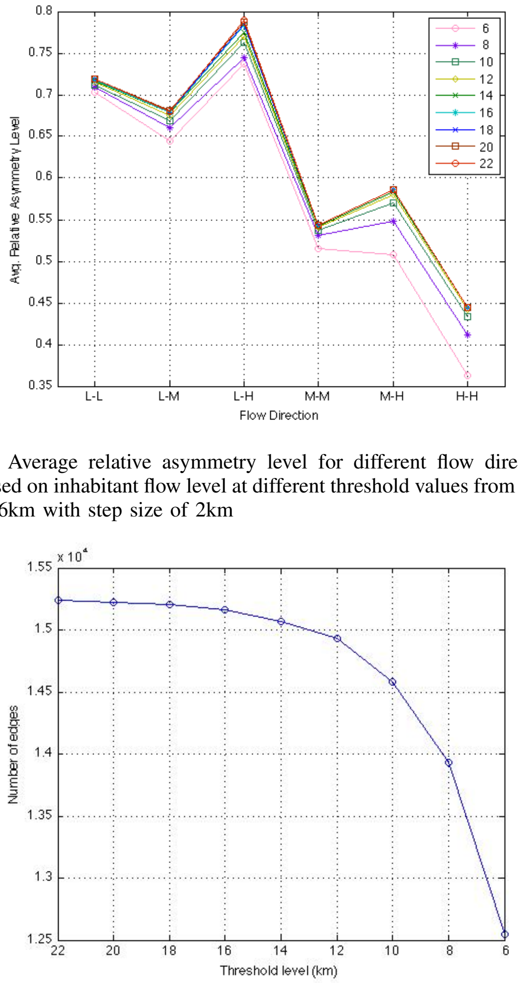
TABLE IV.
AVERAGE RELATIVE ASYMMETRY LEVEL AND THE NUMBER OF FLOWS OF EACH DIRECTION BASED ON POPULATION DENSITY WHERE L=LOW, M=MEDIUM, AND H=HIGH

Group of Cell Pairs	Group Size	Avg. Relative Asymmetry Level
L-L	4,621	0.67
L-M	3,202	0.66
L-H	323	0.66
M-M	4,722	0.53
M-H	1,958	0.56
H-H	449	0.45

value to test the consistency and reliability of the result. In Fig. 13, the average relative asymmetry level for varying threshold value between 22km and 6km. Clearly, the result is consistent and hence it is not influenced by the errors or outliers of the data.

In addition to the correlation analysis of asymmetry level associated with different group of cell pairs based on flow volume that suggests that high asymmetrical flows occur mostly between high and low congested areas, we further investigate the relationship between the asymmetrical flows and population density. Based on 2000 U.S. Census data of Suffolk county's population density, *we find that the flows from and to low populated areas are at high degree of asymmetry* as seen in the results shown in Fig. 14. The details are given in Table IV where groups are defined as follows: High: population $\geq 20,000/\text{km}^2$, Medium: 5,000/km$^2 \leq$ population $< 20,000/\text{km}^2$, and Low: population $< 5,000/\text{km}^2$. Note: each grid cell is assigned with the population density specified by the most covered Voronoi cell.

With land use information obtained from WebGIS [53], we are also able to examine the asymmetry level associated with the flows across different categories of land use. From 33 land use categories, we classify these categories and consider four groups for our analysis: Commercial, Recreational, Residential, and Natural. As shown in Fig. 15, *flows entering and departing commercial areas have high asymmetry levels*. This is consistent



(a) Average relative asymmetry level for different flow directions based on inhabitant flow level at different threshold values from 22km to 6km with step size of 2km

(b) Number of total edges decreases exponentially as threshold value (in km) decreases

Figure 13. Testing the reliability of the result obtained in Fig. 12 by using several threshold values. The result is consistent and hence it is not influenced by the errors or outliers in the data.

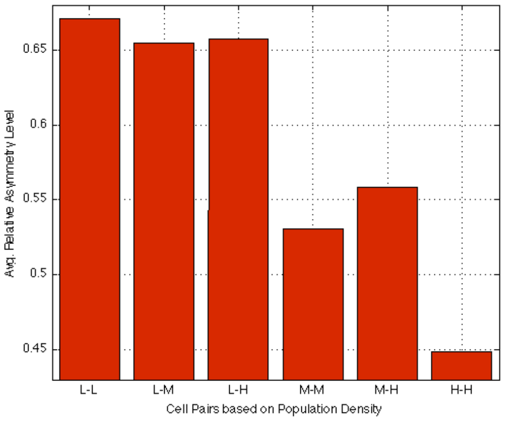


Figure 14. Average relative asymmetry level for different groups of cell pairs based on population density range where L=Low, M=Medium, and H=High.

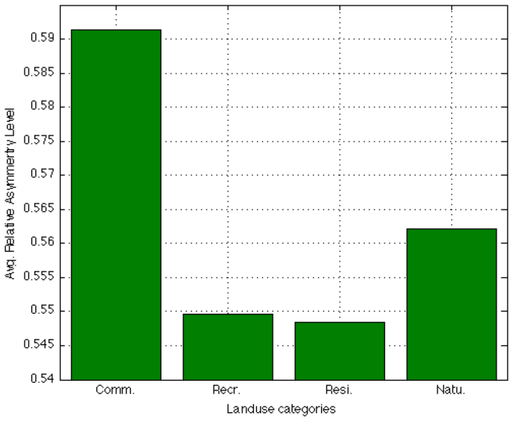


Figure 15. Average relative asymmetry level for different groups of cell pairs associated with different land use categories where Comm. = Commercial, Recr. = Recreational, Resi. = Residential, and Natu. = Natural.

with the previous result obtained from population density that high asymmetrical flows are observed mostly in the cell pairs that are associated with low populated cells (e.g. commercial areas). The flows through residential areas appear to have low asymmetry levels. As we have shown that people would eventually return to their homes after their trips to some destinations (which is reflected by symmetry of within-cell flow), therefore the asymmetry level is expected to be low around the residential areas and hence the result is intuitive.

IV. Applications to O-D Matrix Estimation

The O-D matrix is difficult and often costly to obtain by direct measurements, interviews, or surveys. Using mobile phone data to estimate inhabitant flow and its symmetry associated with different areas or zones can be beneficial to the O-D matrix estimation. Here we outline a number of prospective applications of the results of our study in facilitating O-D matrix estimation as following:

1) As alternative to the maximum entropy - minimum information criterion (one of the most classical approaches in estimating OD-pairs from traffic counts), Bayesian statistical inference has been proposed to combine the reliability of information gathered from the a priori trip matrix (prior information) and the one gathered from link traffic counts [34]. The main feature of this approach is to assign an a priori set of weights for the a priori trip matrix. Thus, this Bayesian updating approach can potentially balance the information of link traffic counts with the flow estimated using mobile phone data and other sources of information, apart from the a priori trip matrix.

2) Generally, the quality of an estimated O-D matrix depends much on the reliability of the input data as well as the number and locations of traffic counting points in the road network. Yang and Zhou [54] formulate a rigorous mathematical framework with the objective of finding the set of link count locations and the minimum number that minimizes the *Maximal Possible Relative Error* (MPRE) i.e. maxmin-risk averse methodology. The developed framework is founded on the following rules: *OD-covering rule*, *Maximum flow fraction rule*, *Maximum flow-intercepting rule*. and *Link independence rule*. Our result on the flow and its symmetry of different divided areas will facilitate Yang and Zhou's approach as a proxy in finding the optimal locations and number of counters according to the developed rules.

3) The potentials of data fusion is envisaged as a methodology to better catch the dynamics and the uncertainty of the real traffic states [52]. Thus, the flow and its symmetry estimated using mobile phone data can be used to give the extra feedback, especially to understand the reliability of data collected from different sources for O-D matrix estimation.

4) The result of this study provides the estimated flow state, which helps at finding automatic procedures for reliable zoning solutions envisaged by Viti [52]. By knowing state of the flow of the areas, one can find in some ways sets of areas that are correlated with one another and on the other hand as much disjoint as possible. This can lead to separated zones with similar or mutually explainable flow (traffic) states.

5) A crucial point in the estimation of an O-D matrix using traffic counts is the assignment technique used: what route(s) in the transport network do trips from zone i to zone j take. The result of our study accommodates the treatment of the congestion effects by estimating the congestion level of different areas such that the technique used for the assignment of traffic on the routes connecting each O-D pair can be decided accordingly: *proportional assignment* is suitable for low level of congestion [55] [6] while *equilibrium assignment* is more realistic approach for high level of congestion [35] [32].

6) As most models assume or require that a target O-D matrix is available [1], the result of our study helps estimating the target O-D matrix (number of travelers attracted to/originating in different different zones), which is typically obtained by a (costly) sample survey or from an old (probably outdated) matrix.

7) Creating a balance in the traffic network is desirable. One aspect of traffic management is to reduce attractiveness of an area either by shifting some

important attractive avenues in it to some other place, or by making other area more competitive in comparison to the present one. If the attractiveness of an area is very high, the public buildings like schools, hospitals, etc. should be planned in the area. On the other hand, if the attractiveness of an area is low, one should increase its attractiveness by planning various important public buildings there to create balance in the transportation network, which reduces unnecessary crowding of vehicles/people and number of accidents [2]. The attractiveness can be estimated using asymmetry of the flow and its geographical location. It reflects in our result as the flows coming to/outing out of central Boston area (with a high attractiveness) have high asymmetry level.

V. Conclusions

Understanding inhabitant mobility is essential for planning and design of city and public transportation. In this research, we carry out a study of the inhabitant flow. We have shown with a large mobile phone data that provides more than two million location traces in Suffolk county, Massachusetts that people begin and end their trips at the same locations (e.g. homes). We have also shown that people travel in an asymmetrical way, which means that they do not always take the completely same routes between home-to-destination and destination-to-home direction. We have shown that these asymmetrical flows account for about 33% of the entire flows, which means that one third of the flows in Suffolk county, MA are imbalanced. In addition, the asymmetrical trips mostly take place while traveling between high and low congested areas such as trips between urban and suburban areas, as well as traveling to and from low populated areas such as commercial areas. To our knowledge, our study is the first report that reveals the asymmetry of inhabitant flows through a large mobile phone data. This study also adds to other researchers' findings on route choice behavior by focusing on the flow of inhabitants over extensive period of time. In addition, the knowledge of commuting flows can be used to facilitate urban planning and design by optimizing the symmetry levels in desired areas to create a better urban system. As our future direction, we will continue to investigate the inhabitant flow and its relationship to the spatial characteristics. The study can also be extended to explore the temporal pattern in the flow as well as the comparisons in flow properties among different cities, counties, and states. Individual traces can also be examined with street networks for analysis of rote choice behavior.

References

[1] T. Abrahamsson. Estimation of origin-destination matrices using traffic counts – a literature survey. *Interim Report, IR-98-021/May, International Institute for Applied Systems Analsyis*, 1998.

[2] A. K. Agrawal, D. Mohan, and R. S. Singh. Traffic planning in a constrained network using entropy maximisation approach. *Journal of the Institution of Engineers. India. Civil Engineering Division*, 85:236–240.

[3] Airsage. Airsage wise technology. *http://www.airsage.com*.

[4] T. S. Azevedo, R. L. Bezerra, C. A. V. Campos, and L. F. M. de Moraes. An analysis of human mobility using real traces. In *WCNC'09: Proceedings of the 2009 IEEE conference on Wireless Communications & Networking Conference*, pages 2390–2395, Piscataway, NJ, USA, 2009. IEEE Press.

[5] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. In *Proc. Nat. Acad. Sci. 101*, pages 37–47, 2004.

[6] M. Bell. Log-linear models for the estimation of origin-destination matrices from traffic counts. In *Proc. of the Ninth International Symposium on Transportation and Traffic Theory*, 1984.

[7] P. Bovy and E. Stern. Route choice: Wayfinding in transport networks. *Kluwer Academic Publishers*, 1990.

[8] F. Calabrese, F. C. Pereira, G. D. Lorenzo, and L. Liu. The geography of taste: analyzing cell-phone mobility and social events. In *Proceedings of IEEE Inter. Conf. on Pervasive Computing (PerComp)*, 2010.

[9] J. Candia, M. C. Gonzalez, P. Wang, T. Schoenharl, G. Madey, and A. Barabasi. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):1–16, 2008.

[10] R. Conroy Dalton. The secret is to follow your nose: Route path selection and angularity. *Environment and Behavior*, 35:107–131, 2003.

[11] R. Dial. A probabilistic multipath traffic assignment algorithm which obviates path enumeration. *Transportation Research*, 5(2):83–111, 1971.

[12] M. Duckham and L. Kulik. Simplest paths: Automated route selection for navigation. In *COSIT 2003, Lecture Notes in Computer Science*, volume 2825, pages 169–185, 2003.

[13] N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, May 2006.

[14] P. Gipps. Simulation of pedestrian traffic in buildings. *Schriftenreihe des Instituts fuer Verkehrswesen, University of Karlsruhe*, 1986.

[15] R. G. Golledge. Path selection and route preference in human navigation: A progress report. *Transportation Research*, B 38:169–190, 2004.

[16] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[17] Y. Guy. Pedestrian route choice in central jerusalem. department of geography. *Ben-Gurion University of The Negev, Beer Sheva (in Hebrew)*, 1987.

[18] H. Hamacher and S. Tjandra. Mathematical modeling of evacuation problems: A state of the art. In *Proceedings of the Pedestrian and Evacuation Dynamics*, pages 59–74. Springer, Berlin, 2001.

[19] D. Helbing. Traffic dynamics: New physical modeling concepts. *Springer-Verlag, Berlin (in German)*, 1997.

[20] M. Hill. Spatial structure and decision-making of pedestrian route selection through an urban environment. *Ph.D. Thesis, University Microfilms International*, 1982.

[21] B. Hillier and J. Hanson. *The Social Logic of Space*. Cambridge University Press, Cambridge, 1984.

[22] H. Hochmair and A. U. Frank. Influence of estimation errors on wayfinding decisions in unknown street networks analyzing the least-angle strategy. *Spatial Cognition and Computation*, 2:283–313, 2002.

[23] S. Hoodendoorn and P. Bovy. Pedestrian route-choice and activity scheduling theory and models. In *Spatial Information Theory A Theoretical Basis for GISTransportation Research*, pages 207–222, 1995.

[24] R. Hughes. A continuum theory for the flow of pedestrians. *Transportation Research*, B 36(6):507–535, 2002.

[25] W. S. Jung, F. Wang, and H. E. Stanley. Gravity model in the korean highway. *Europhys. Lett.*, 81(48005), 2008.

[26] Z. G. K. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Reading, MA: Addison-Wesley, 1949.

[27] G. Krings, F. Calabrese, C. Ratti, and V. D. Blondel. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, pages 1–8, 2009.

[28] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong. Slaw: A mobility model for human walks. In *Proceedings of the 28th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, Rio de Janeiro, Brazil, April 2009. IEEE.

[29] N. Limtanakool, M. Dijst, and T. Schwanen. Development of hierarchy in the dutch urban system on the basis of flows. In *Colloquium Vervoersplanologisch Speurwerk 2005: Duurzame mobiliteit: hot or not?*, pages 21–40, 2005.

[30] N. Limtanakool, M. Dijst, and T. Schwanen. A theoretical framework and methodology for characterising national urban systems on the basis of flows of people: Empirical evidence for france and germany. *Urban Studies*, 44(11):2123–2145, 2007.

[31] N. Limtanakool, M. Dijst, and T. Schwanen. Developments in the dutch urban system on the basis of flows. *Regional Studies*, 42(2):179–196, 2009.

[32] L. L.J. and F. K. Selection of a trip table which reproduces observed link flows. *Transportation Research 16B*, 1982.

[33] G. Lovas. Modeling and simulation of pedestrian traffic flow. *Transportation Research*, B 28(6):429–443, 1994.

[34] M. Maher. Inferences on trip matrices from observations on link volumes: a bayesian statistical approach. *Transportation Research Part B*, 17B(6):435–447, 1983.

[35] S. Nguyen. Estimation of an od matrix from network data: A network equilibrium approach. *University of Montreal, Quebec, Canada*, (60), 1977.

[36] T. A. S. Nielsena and H. H. Hovgese. Exploratory mapping of commuter flows in england and wales. *Journal of Transport Geography*, 16(2):90–99, 2008.

[37] S. Phithakkitnukoon and R. Dantu. UNT mobile phone communication dataset. *http://nsl.unt.edu/santi/data_desc.pdf*, 2008.

[38] S. Phithakkitnukoon and R. Dantu. Mobile social group sizes and scaling ratio. *Springer: AI & Society*, 2009.

[39] S. Phithakkitnukoon and R. Dantu. Mobile social closeness and similarity in calling patterns. In *IEEE Conference on Consumer Communications & Networking Conference (CCNC 2010)Special Session on Social Networking (SocNets)*, 2010.

[40] S. Phithakkitnukoon, T. Horanont, G. D. Lorenzo, R. Shibasaki, and C. Ratti. Activity-aware map: Identifying human daily activity pattern using mobile phone data. In *Proc. of Inter. Conf. on Pattern Recognition (ICPR 2010), Workshop on Human Behavior Understanding (HBU)*, 2010.

[41] M. Raento. Context project. *http://www.cs.helsinki.fi/group/context/data/*, 2008.

[42] J. Rich, S. L. Mabit, and O. A. Nielsen. Route choice model for copenhagen: A data-driven choice set generation approach based on gps data. In *http://transpor2.epfl.ch/tristan/FullPapers/068Rich.pdf.*

[43] E. Sadalla, W. Burroughs, and L. Staplin. Reference points in spatial cognition. *Journal of ExperimentalPsychology: Human Learning and Memory*, 5:516–528, 1980.

[44] P. Senevarante and J. Morall. Analysis of factors affecting the choice of route of pedestrians. *Transportation Planning and Technology*, 10:147–159, 1986.

[45] Skyhook. Skyhook wireless. *http://www.skyhookwireless.com/.*

[46] C. Song, Z. Qu, N. Blumm, and A.-L. Barabsi. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[47] J. Teklenburg, H. Timmermans, and A. Borgers. Changes in urban layout and pedestrian flows. In *Proceedings of Seminar A PTRC European Transport Forum*, volume P363, pages 97–108, 1993.

[48] A. Turner. Angular analysis. In *Proceedings of the 3rd International Symposium on Space Syntax*, pages 30.1–30.11, Georgia Institute of Technology, Atlanta, 2001.

[49] A. Turnerand and N. Dalton. A simplified route choice model using the shortest angular path assumption. In *www.geocomputation.org/2005/Turner.pdff.*

[50] J. R. van Eck and D. Snellen. Is the randstad a city network? evidence from commuting patterns. In *European Transport Conference 2006*, 2006.

[51] N. Verlander. Pedestrian route choice: An empirical study. In *Proceedings of Senimar F of the PTRC European Transport Forum*, volume P415, pages 39–49, 1997.

[52] F. Viti. State-of-art of o-d matrix estimation problems based on traffic counts and its inverse network location problem: perspectives for application and future developments. *Working paper*, 2008.

[53] WebGIS. Geographic information systems resource. *http://www.webgis.com/index.html.*

[54] H. Yang and J. Zhou. Optimal traffic counting locations for origin-destination matrix estimation. *Transportation Research Part B*, 32B(2):108–126, 1998.

[55] H. V. Zuylen and L. Willumsen. The most likely trip matrix estimated from traffic counts. *Transportation Research 14B*, 1980.

Santi Phithakkitnukoon is a postdoctoral research fellow at the SENSEable City Laboratory at MIT. He received his B.S. and M.S. degrees in Electrical Engineering from Southern Methodist University, Dallas, Texas, USA in 2003 and 2005, respectively. He received his Ph.D. degree in Computer Science and Engineering from the University of North Texas, Denton, Texas, USA in 2009. His research interests include machine learning and its applications in urban computing, context-aware computing, and mobile/online social analysis.

Professor Carlo Ratti is the director of the SENSEable City Laboratory at MIT, Cambridge, Massachusetts, and an adjunct professor at Queensland University of Technology, Brisbane, Australia. He's also a founding partner and director of the design firm carloratti-associati – Walter Nicolino & Carlo Ratti. Carlo has a Ph.D. from the University of Cambridge and he is a member of the Ordine degli Ingegneri di Torino and the Association des Anciens Elèves de l'École Nationale des Ponts et Chaussées.