From Taxi GPS Traces to Social and Community Dynamics: A Survey

PABLO SAMUEL CASTRO and DAQING ZHANG, Institut Mines-TELECOM/TELECOM SudParis; 9, rue Charles Fourier; 91011 Evry Cedex, France
CHAO CHEN, Institut Mines-TELECOM/TELECOM SudParis; 9, rue Charles Fourier; 91011 Evry Cedex, France; and Universite Pierre et Marie Curie; 4, place Jussieu; 75005 Paris, France
SHIJIAN LI and GANG PAN, Zhejiang University; Hangzhou, 310027, P.R., China

Vehicles equipped with GPS localizers are an important sensory device for examining people's movements and activities. Taxis equipped with GPS localizers serve the transportation needs of a large number of people driven by diverse needs; their traces can tell us where passengers were picked up and dropped off, which route was taken, and what steps the driver took to find a new passenger. In this article, we provide an exhaustive survey of the work on mining these traces. We first provide a formalization of the data sets, along with an overview of different mechanisms for preprocessing the data. We then classify the existing work into three main categories: social dynamics, traffic dynamics and operational dynamics. Social dynamics refers to the study of the collective behaviour of a city's population, based on their observed movements; Traffic dynamics studies the resulting flow of the movement through the road network; Operational dynamics refers to the study and analysis of taxi driver's *modus operandi*. We discuss the different problems currently being researched, the various approaches proposed, and suggest new avenues of research. Finally, we present a historical overview of the research work in this field and discuss which areas hold most promise for future research.

Categories and Subject Descriptors: A.1 [**General Literature**]: Survey

General Terms: Design, Experimentation, Human factors

Additional Key Words and Phrases: Taxi GPS, urban computing, smart cities

ACM Reference Format:
Castro, P. S., Zhang, D., Chen, C., Li, S., and Pan, G. 2013. From taxi GPS traces to social and community dynamics: A survey. ACM Comput. Surv. 46, 2, Article 17 (November 2013), 34 pages.
DOI: http://dx.doi.org/10.1145/2543581.2543584

1. INTRODUCTION

The past decade has seen a dramatic increase in the number of personal devices such as cell phones, portable computers, and GPS localizers. These devices leave digital footprints of their user's activities, which are a reflection of the economical and societal interactions of a community. Using techniques from a wide range of fields, researchers have sought to extract communal behaviours and intelligence from this data to obtain a better understanding of the underlying dynamics of an individual, community, or city [Martino et al. 2010; Zhang et al. 2011a]. This understanding can enable many innovative applications in city planning, traffic management, disease containment, and

Authors' address: P. S. Castro and D. Zhang, Institut Telecom SudParis, 9 rue Charles Fourier, EVRY Cedex, 91011, France.
This research was supported by the Institut TELECOM "Futur et Ruptures" Program, AQUEDUC–Paris-Region SYSTEM@TIC Smart City Program, and the High-Tech Program of China (863) (No. 2011AA010104).

DOI: http://dx.doi.org/10.1145/2543581.2543584

green computing, to name a few. Each different source of data reveals different aspects of the underlying community dynamics, depending on the type of digital footprint.

Vehicles equipped with GPS localizers provide an important type of footprint, given that public and private vehicles are the main transportation means for a city's population. People use vehicles for commuting to and from work, for regular and "irregular" chores, and for leisure activities. By careful analysis of the observed movements of a population, researchers strive to better understand the demographics of a city, the distribution of services around a city, the effectiveness of the various transportation networks, the dynamics of traffic conditions, and the different driving behaviours.

Public transportation vehicles equipped with GPS localizers offer rather predictable data, as the vehicles in question follow fixed routes under a specified schedule. Similarly, private vehicles equipped with GPS localizers are usually restricted to one user and follow fairly predictable routes (i.e., to and from work). On the other hand, taxis equipped with GPS localizers serve the transportation needs of a large number of people driven by diverse needs and are not constrained to a prespecified route. The GPS traces of a taxi can tell us, with a fair amount of precision, where passengers were picked up, where they were dropped off, which route was taken, and what steps the driver took to find a new passenger. By virtue of the diversity of passengers served, as well as their continual operation, a taxi GPS traces offer a rich and detailed glimpse into the motivations, behaviours, and resulting dynamics of a city's mobile population.

In this article, we aim to survey existing research on mining taxi GPS traces as well as suggest some interesting avenues for future research. Through thorough analysis, we group the existing body of work into three categories: social dynamics, traffic dynamics, and operational dynamics. We define *social dynamics* as the work studying the collective behaviour of a city's population, as observed by their movement in the city. Researchers in this field are interested in where people are going throughout the day, the "hottest" spots around a city, the "functions" of these hotspots, how strongly connected are different areas of the city, and so forth. These movements are motivated by diverse needs and influenced by external factors (e.g., weather and traffic). A deep understanding of social dynamics is essential for the management, design, maintenance, and advancement of a city's infrastructure, and is studied by many different fields. Giannotti et al. [2011] list a number of questions of particular interest to mobility agencies, which are a part of the discussion we present herein.

Governed by their underlying desires or needs, people will move around the city mainly through the road network. Whilst social dynamics aims to understand people's movementpatterns, *traffic dynamics* studies the resulting *flow* of the population through the city's road network. Most of this work is generally aimed at predicting traffic conditions and can be useful for providing real-time traffic indicators and travel time estimation for drivers.

Operational dynamics refers to the general study and analysis of taxi drivers' modus operandi. The aim is to be able to *learn* from taxi drivers' excellent knowledge of the city, as well as to *detect* abnormal behaviours. The last two categories used mainly the end points of a taxi's trajectory (i.e., the pickup and dropoff locations); in the study of operational dynamics, researchers make use of full trajectories, as the routes taken by drivers are of utmost importance. Researchers have mined these trajectories in order to predict future trajectories, suggest strategies for quickly finding new passengers, and suggesting navigational routes for reaching a destination quickly. Additionally, new trajectories can be compared against a large collection of historical trajectories to automatically detect abnormal behaviour.

It is worth stressing that although we have classified the various works into three categories, they are by no means independent of each other. Most papers, despite focusing on one particular category, deal with subjects relating to the other categories.

Table I. Fields for a GPS Entry with a Sample

Taxi ID	Longitude	Latitude	Speed (km/h)	Bearing	Occupied flag	Year	Month	Day	Hour	Minute	Second
10429	120.214134	30.212818	70.38	240.00	1	2010	2	7	17	40	46

As an example, consider the work of Li et al. [2011b] on discovering passenger finding strategies. Although passenger-finding strategies fall within the Operational Dynamics category, their method requires the extraction of 'hotspots, which falls under the Social Dynamics category; additionally, the information produced by their method gives some insight into the traffic conditions around different parts of the city at different times, which would fall under the Traffic Dynamics category. This dependence across categories is something we encounter in most of the papers we surveyed. It is thus difficult to provide an exhaustive survey of work on mining taxi GPS traces without surveying all three categories in conjunction.

This article is organized as follows. In Section 2, we begin by providing a formal definition of the dataset, along with a survey of existing work in preprocessing this data. In Section 3, we review the work relating to social dynamics; in Section 4, we analyze the work relating to traffic dynamics; and in Section 5, we survey the papers focusing on navigational strategies. In Section 6, we present some deployed systems making use of taxi GPS information. We provide a brief historical overview of the development of this type of work in Section 7, and we conclude the paper in Section 8. In surveying these papers, we might reference certain techniques or tools with which some readers may not be familiar. As providing a description for each of these would disrupt the flow of the presentation and drastically increase the size of the article, we recommend that interested readers follow the references provided to seek more detailed information.

2. DATA PREPARATION

In this section, we define the taxi GPS dataset that will be the basis of the research surveyed. For concreteness, we present the dataset used by our research group, but the differences amongst different datasets are of little significance (as seen in Wang et al. [2011],where three different datasets from three cities in China are presented). As previously mentioned, the raw data must be prepared in order for it to be suitable for the work discussed in the sequel. We survey a number of papers addressing the data preparation problem, including Edelkamp and Schrödl [2003], Schroedl et al. [2004], Cao and Krumm [2009], Haklay and Weber [2008], Qi et al. [2011], Gonzalez et al. [2007], Yuan and Zheng [2010], and Zheng et al. [2011].

2.1. Dataset Description

We consider situations where the GPS records of a large number of taxis in a city are routinely saved to a log file over a number of months, resulting in a very large dataset. Our dataset was obtained from around 7,000 taxis in Hangzhou, China, over a period of 12 months, at a sampling rate of approximately once per minute (for each taxi), resulting in more than a billion GPS entries. Table I lists the fields for each GPS record, along with a sample entry.

Henceforth, we shall use the variable σ to represent a GPS entry, with fields (in the same order as in Table I) $\{id, lon, lat, speed, bear, occ, year, month, day, hour, minute, second\}$. Let Σ be the set of all GPS entries.

Although the data for each taxi can be considered as a continuous stream throughout the full year, we split it up into trajectories, based on whether the taxi is occupied or unoccupied (the occupancy flag is consistent within each trajectory). Employing the

taxi occupancy flag can yield an advantage over other methods for splitting trajectories, such as used by Giannotti et al. [2011], where they split GPS trajectories of regular vehicles by detecting more than 30-minute "pauses" in the transmission of logs.

Definition 2.1. A GPS trajectory τ of length n consists of a series of GPS entries $\langle \sigma^t \dots \sigma^{t+n} \rangle$ such that for all $t \le i, j \le t+n$, $\sigma^i_{occ} = \sigma^j_{occ}$. A trajectory τ is an occupied trajectory if for all i, $\sigma^i_{occ} = 1$, and a vacant trajectory otherwise.

We can then see that the data for a single taxi consists of a series of alternating occupied and vacant trajectories. The end points of the occupied trajectories yield some information that will be useful for many of the papers presented.

Definition 2.2. Given an occupied trajectory $\tau = \langle \sigma^1 \dots \sigma^n \rangle$, we define σ^1 as the pickup point and σ^n as the dropoff point.

Let $\chi \subset \mathbb{R}^2$ be the set of distinct latitude/longitude positions in Σ, and define the boundary latitude and longitudes as:

$$lon_{min} = \min_{\sigma \in \Sigma} \sigma_{lon} \qquad lat_{min} = \min_{\sigma \in \Sigma} \sigma_{lat}$$

$$lon_{max} = \max_{\sigma \in \Sigma} \sigma_{lon} \qquad lat_{max} = \max_{\sigma \in \Sigma} \sigma_{lat}$$

2.2. City Decomposition

The "terrain" of a city is a continuous two-dimensional area (i.e., a subset of $\mathbb{R}^2$), which is difficult to work with. It is more practical to decompose the city into separate (usually disjoint) areas and work with this decomposition. Let $\mathbb{D} \subseteq \mathbb{R}^2$ be the area covering all of the GPS entries. Specifically,

$$\mathbb{D} = [lon_{min}, lon_{max}] \times [lat_{min}, lat_{max}]$$

Definition 2.3. A city decomposition is a pair (Ψ, ψ), where Ψ is a finite set of "areas" and $\psi : D \to \Psi$ is a membership function, where $D \subseteq \mathbb{R}^2$, mapping any location $(x, y) \in D$ to an area given by $\psi(x, y) \in \Psi$.

We will often make use of certain functions to characterize areas based on some statistical criterion. We formally define these functions next and present instances of them throughout the article.

Definition 2.4. A characterization function $f : \mathcal{P}(\mathbb{R}^2) \to \mathbb{R}^d$ maps any set of points (specified by latitude/longitude) to a characteristic vector in $\mathbb{R}^d$ (given a set X, we denote its power set by $\mathcal{P}(X)$).

Given a decomposition (Ψ, ψ), we can apply a characterization function to an area A by $f(\psi^{-1}(A))$. The usefulness of a characterization function highly depends on the decomposition used, and we should require these functions to be consistent with the decomposition, as defined next.

Definition 2.5. A characterization function f is consistent with a decomposition (Ψ, ψ) if, given an area $A \in \Psi$ split into two "subareas" $B, C \subseteq \mathbb{D}$ such that $B \cup C = \psi^{-1}(A)$ and $|B| \approx |C|$, then $f(\psi^{-1}(A)) \approx f(B) \approx f(C)$.

The idea behind consistency is that the characterization function in question should behave uniformly throughout each area, and there should be little loss of information introduced by the decomposition. Note that the number of characterization functions that are consistent with a decomposition is a good indication of the quality of the decomposition. In a sense, one can use characterization functions to measure the *homogeneity* of the areas in a decomposition.

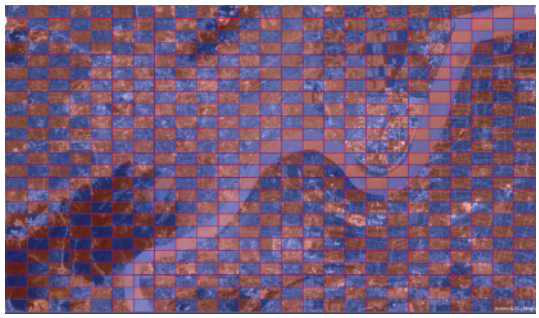

Fig. 1. A 25 × 25 grid decomposition of Hangzhou. Each area is roughly 1km^2. ©Google (2011).

The purpose of decomposition is to render the continuous space $\mathbb{D}$ manageable for analysis. A very simple approach could be to use χ directly (i.e., one element for every distinct lat/lon point), which suffers from no information loss. However, the latitude and longitude readings obtained from the GPS devices are quite sensitive, so there may be small variations even when the vehicle is not moving. Given that we have more than a billion GPS entries, the size of χ would be prohibitively large. With this in mind, we will review the two most popular approaches to city decomposition: splitting the city into equally sized grids and using a digital map.

2.2.1. Grid Decomposition. A popular and simple approach to decomposing a city is to split up the city's areas into equal-sized grids. This method has been used in many papers, including Girardin et al. [2008], Phithakkitnukoon et al. [2010], Huang et al. [2010], Liu et al. [2010b], and Zhang et al. [2011b]. The advantage of this method is that it is very simple and easy to implement, and one can directly determine the size of the resulting decomposition by changing the size of the grids. Figure 1 presents an example of this type of decomposition.

An $(m_1 \times m_2)$-grid decomposition consists of an $m_1 \times m_2$ matrix of grid cells $\{(i, j)\}_{1 \le i \le m_1, 1 \le j \le m_2}$, where the width of each grid cell is $\Psi_w = \frac{lon_{max} - lon_{min}}{m_2}$, and the height of each grid cell is $\Psi_h = \frac{lat_{max} - lat_{min}}{m_1}$. The function ψ is defined as follows, for any $(x, y) \in \mathbb{D}$:

$$\psi(x, y) = \left(\arg \min_{1 \le i \le m_1} \{x < lat_{min} + i * \Psi_h\}, \arg \min_{1 \le j \le m_2} \{y < lon_{min} + j * \Psi_w\} \right)$$

This type of decomposition is usually consistent with characterization functions that are based on *averages*, such as average speed, average number of pickup/dropoffs, and expected average emissions. This is because these measurements are *normalized*, effectively removing the effect of the size of the area. We stress that this *usually* holds but is not always the case. Consider, for instance, a grid cell that has a section of a river on its left half and a part of the road network on the right half. The left subarea contains no streets, so the value assigned by any characterization function to the left subarea will be drastically different to that assigned to the original area. This problem can be somewhat mitigated by using smaller grids, but because of the static nature of this decomposition, one cannot guarantee that all of the grids will be homogeneous with respect to the desired characterization functions.

Nonnormalized characterization functions (e.g., max/min speed, total number of pick-up/dropoffs) will, for the most part, not be consistent with this type of decomposition, as the value of the characterization functions will depend greatly on the size and location of each grid cell.

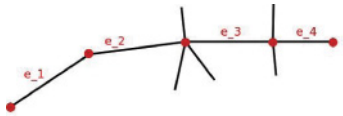

Fig. 2. A street split into various edges.

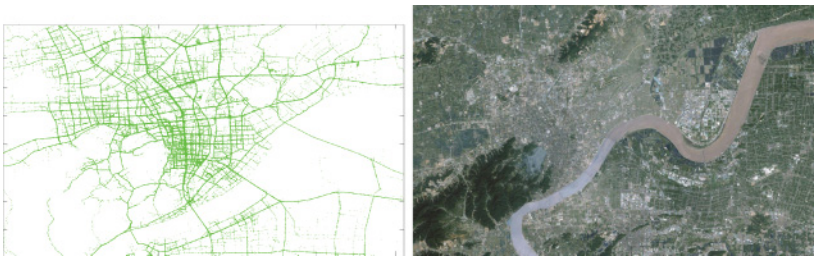

Fig. 3. Left: Traces from 20 taxis over 1 month; Right: Digital road network drawn over raw traces, with some important areas highlighted. ©Google (2011).

The advantages of this type of decomposition is that it is simple and easy to implement. It splits the city into disjoint areas that are easy to visually inspect, which is of particular interest for qualitative results. As we will discuss later, one can perform further processing on the grid cells (e.g., clustering and removal of "empty" areas), yielding a decomposition that is more adapted to the dataset.

One major disadvantage of this decomposition method is that the number and size of the grid cells is determined independently of the dataset. This results in busy areas decomposed in the same manner as areas that have very little activity. For instance, although 1km^2 areas may be suitable for suburban or rural areas, it is most likely too coarse for the downtown area. On the other hand, using very small grids will result in a large Ψ and will be unnecessarily small for areas with little activity (e.g., the mountainous area in the bottom left of Figure 1, where there are no streets). Another disadvantage is that many road segments are inevitably grouped together in the same cell. This can be problematic if one is interested in routing information (e.g., common routes taken by commuters) or road congestion estimation.

2.2.2. Digital Map. If a digital map of the city is available, one can map each GPS entry onto a point on the digital map, with the main elements defined next.

Definition 2.6. A digital map is a graph (V, E), where V is a set of vertices and E is a set of edges. Each edge $e \in E$ has the following fields: two endpoint vertices (whose coordinates correspond to lat/lon values) e_{v_1} and e_{v_2}; length e_{length} and bearing e_{brng}.[1]

A street can consist of multiple edges, as shown in Figure 2. Note that it is split into segments wherever there is an intersection and/or a change in bearing.

In many cases, a digital map of the city is not readily available. Given the long time span and large number of taxis available in our dataset, the city's road network becomes apparent by simply plotting all of the GPS entries. In the left-hand panel of Figure 3, we display a sample of the plot obtained from 20 taxis over 1 month, around the downtown area of Hangzhou. We only plot the pickup and dropoff points, as we found that including full trajectories clutters the plot. A simple (but arduous) option is to "draw" the edges and construct a digital map manually. In the right-hand panel

[1]Note that the bearing is the direction from e_{v_1} to e_{v_2}; simply substract this bearing from 360 to obtain the bearing from e_{v_2} to e_{v_1}.

of Figure 3, we display the resulting digital map. It consists of 2,003 edges and 1,585 vertices.

Drawing in a digital map can be an arduous process, so many researchers have investigated constructing digital maps automatically from digital traces. Edelkamp and Schrödl [2003] construct a digital map representation (including lanes) by matching GPS traces to an existing base map and/or inferring road segments by clustering traces and applying further statistical refinements. Schroedl et al. [2004] infer road segments and intersections by clustering together the trajectories. Worrall and Nebot [2007] construct a compact digital map from a fleet of GPS-equipped mining vehicles by clustering according to their location and heading, and then compressing these many segments into a set of lines and arcs (according to curvature). Cao and Krumm [2009] automatically build a digital map from raw GPS traces by "pulling" together traces into paths, which are then clustered to create a graph of nodes and edges. Fathi and Krumm [2010] use a localized shape descriptor to represent the distribution of GPS traces around a point in order to automatically detect road intersections from the GPS traces. Biagoni and Eriksson [2012] provide a comprehensive survey and comparison of different approaches to map generation from GPS traces.

One disadvantage of using digital maps as we have described them here is that they lack some crucial information, such as width of each road segment, the number of lanes, and orientation of each lane. Gathering this type of information manually would be quite expensive. There have been a number of papers that aim to use GPS traces to automatically enhance an existing road network. After the initial digital map is constructed, Schroedl et al. [2004] establish lane positions by clustering offsets from a previously found segment centerline. Similarly, Rogers et al. [1999] model roads via their centerline and lanes as fixed offsets from this centerline. Chen and Krumm [2010] use a modified Gaussian mixture model for estimating the distribution of GPS traces across multiple lanes. OpenStreetMap combines GPS traces, satellite images, and hand-labelled information to produce a very rich digital map [Haklay and Weber 2008]. Alvares et al. [2007] propose finding "stops" and "moves" common to most taxis to enrich existing GPS trajectories with "landmarks."

Given a digital map (V, E), one can map each GPS entry to a point on one of the edges from E. Simply mapping to the closest edge can suffer from different types of errors, such as mapping to a road network that is perpendicular to the true road network. There are a number of approaches to map matching that take contextual information into consideration, such as edit distance [Yin and Wolfson 2004] and Fréchet distance [Alt et al. 2003; Brakatsoulas et al. 2005]. The previous approaches map trajectories *globally*, that is, only complete trajectories are considered. This renders these approaches unsuitable for long trajectories, as the computational expense becomes too large. Some papers have proposed matching trajectories on a *local* basis, using contextual information such as distance and orientation [Greenfeld 2002; Li et al. 2007], spatial context, and speed information [Lou et al. 2009], or the influence of neighbouring GPS points [Yuan et al. 2010]; Chawathe [2007] proposes assigning a confidence score to different segments of the trajectory and then proceeds to sequentially match the different segments, beginning with those with the highest confidence score. Liu et al. [2012a] propose first pruning a set of trajectories by using speed and orientation, then clustering the remaining segments using distance and orientation, and finally using B-spline fitting [Schroedl et al. 2004] to fit the clustered traces onto road segments.

A GPS entry σ that has been mapped onto a digital map will result in a new entry $\bar{\sigma}$. This new entry is identical to σ, except the latitude and longitude may be different (so that the point "sits on" the digital map), and it is augmented with the fields *edge* and *dir*. The field *edge* represents what edge $e \in E$ it was mapped to, and *dir* represents what street direction it is following (1 if going from e_{v_1} to e_{v_2} and -1 if going in the

other direction). Let $\bar{\Sigma}$ be the set of all mapped GPS entries. Mapped GPS trajectories $\bar{\tau}$ can be defined in the same way as in Definition 2.1, but using entries $\bar{\sigma} \in \bar{\Sigma}$.

Since trajectories will be used for routing purposes further on, we will be interested in edge and/or vertex trajectories, as defined next.

Definition 2.7. An edge trajectory $\varepsilon \in E^*$ consists of a sequence of edges $\langle e^1, e^2, \ldots, e^n \rangle$ such that all adjacent edges are connected: for all $1 \leq i < n$, $\{e^i_{v_1}, e^i_{v_2}\} \cap \{e^{i+1}_{v_1}, e^{i+1}_{v_2}\} \neq \emptyset$.

Definition 2.8. A vertex trajectory $\omega \in V^*$ consists of a sequence of vertices $\langle v^1, v^2, \ldots, v^n \rangle$ such that for all adjacent vertices v^i and v^{i+1}, there exists an edge $e \in E$ such that $e_{v_1} = v^i$ and $e_{v_2} = v^{i+1}$, or vice versa.

We will refer to the ith edge in an edge trajectory ε via $\varepsilon(i)$ and similarly for vertex trajectories. Note that one can extract a valid vertex trajectory from a valid edge trajectory, and vice versa. However, due to low sampling rates, the mapped GPS trajectories may not produce valid edge nor vertex trajectories, as there may be "gaps" (defined next) between GPS entries, or the mapped edges may not have consistent directions.

Definition 2.9. A gap exists between two mapped GPS entries $\bar{\sigma}^1$ and $\bar{\sigma}^2$ if the two edges do not meet: $\{e^1_{v_1} \cup e^1_{v_2}\} \cap \{e^2_{v_1} \cup e^2_{v_2}\} = \emptyset$, where $e^1 = \bar{\sigma}^1_{edge}$ and $e^2 = \bar{\sigma}^2_{edge}$.

The conditions for valid edge and vertex trajectories can be used as guides for choosing the edges to which the GPS entries should map. After computing a mapped GPS trajectory, it is also necessary to complete the gaps to obtain valid edge and vertex trajectories. The way to complete these gaps is unclear, but a simple approach is to compute the shortest path between the two entries with a gap.

We can define digital map decompositions by the pair (Ψ_{DM}, ψ_{DM}), where $\Psi_{DM} = E$ and the membership function ψ_{DM} depends on the map-matching algorithm being used.[2]

This type of decomposition is consistent with the same type of characterization functions that are consistent with the grid decomposition, and because of the (usually) small length of road segments, it overcomes many of the problems present with grid decomposition. The size of the road segments renders this type of decomposition for the most part consistent with nonnormalized characterization functions such as max/min speed, number of pickup/dropoffs, and so forth. Obviously, characterization functions that depend on enhanced information such as lane orientation will not be consistent with the decomposition if the decomposition does not contain that type of information.

Since taxis are restricted to navigating through the road network, a digital map is nearly optimal in terms of overall consistency and size. As will be discussed next, this type of decomposition can be further refined by clustering road segments or splitting based on certain types of vertices or edges. This effectively combines the advantages of both grid and digital map decomposition. The main disadvantage of this type of decomposition is that it relies on a pre-existing digital map, which may not always be available. Some methods were mentioned for their automatic construction, but the quality or usefulness of the resulting digital map is not clear.

2.3. Clustering Areas

The areas resulting from the decompositions mentioned earlier can often benefit from further clustering. This is especially true for the grid-based decomposition, as there may be many "empty" areas, as well as certain adjacent areas that exhibit the same type of behaviour.

[2]We may also have situations where $\Psi_{DM} = V$ or $\Psi_{DM} = V \times E$.

2.3.1. TimeFeature Clustering. The different areas can be characterized by averaging a set of characterization functions over a fixed time window T (such as 24 hours or a number of days or months). This time window is split into d disjoint time "bins" $\{b_1, b_2, \ldots, b_d\}$, where each b_j corresponds to a time interval (e.g., 15 minutes, an hour, 12 hours, a day). Let $\mathcal{B}_T$ be the set of bins corresponding to time period T. The time units used to define the bins depends on the level of granularity specified.

Assume that we have a set of characterization functions $\{c_1, c_2, \ldots, c_n\}$ (see Definition 2.5) in which we are interested, where for any $1 \leq i \leq n$, $c_i : \Psi \times \mathcal{B}_T \to \mathbb{R}$ maps an area and time bin to a real value (e.g., average number of pickup/dropoffs, average speed). Then, for any area $A \in \Psi$, each characterization function c_i gives rise to a vector: $\mathbf{x}^i = (x_1^i, x_2^i, \ldots, x_d^i)$, where $x_j^i = c_i(A, b_j)$, resulting in a collection of vectors $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\}$. We denote this collection the TimeFeature of area A.

One can define a distance function that measures the similarity between two TimeFeatures. This distance function can then be used with a clustering algorithm (such as k-means clustering) to group together similar areas. We will examine two works that use TimeFeature clustering and present some results on our dataset.

Froehlich et al. [2009] used this type of clustering to perform a spatiotemporal analysis of a city based on a shared bicycling system. They use a 24-hour window split into 288 bins, corresponding to 5-minute intervals. They average the values of different characterization functions over the 24-hour window, such as the number of available bicycles, a measure of the activity of the bicycle station, and number of checked out bicycles. They use a metric based on Dynamic Time Warping [Sakoe and Chiba 1990] to allow for temporal pattern shifts between feature vectors. Given these feature vectors and distance function, dendogram clustering [Duda et al. 2001] was used to group together similar bicycle stations.

Qi et al. [2011] use TimeFeature clustering to uncover the different social functions of different regions of a city, using varying windows of 1 week, 1 month, and 1 year split into days (coarse grained) or "quarter" days (fine grained). They use the number of taxi dropoffs at each area as the characterization function. The resulting TimeFeatures are normalized, and the "cosine distance" function is used, which is one minus the cosine of the included angle between two TimeFeatures (considered as vectors in Euclidean space). Finally, a simple agglomerative clustering method [Xu and Wunsch 2005] is used to group similar locations. Principal Component Analysis (PCA) [Jolliffe 1986] was used to reduce the dimensionality of the feature vectors.

We have used this type of clustering with our dataset for Hangzhou. We split the map into a 200×100 matrix of grid cells, each covering around 250m^2. We consideredonly weekdays and used a 24-hour window split into 96 bins, corresponding to 15-minute intervals. We used the dropoff frequency as the characterization function and performed k-means clustering with the cosine distance function, setting $k = 5$ clusters. Figure 4 displays the resulting clusters for grid cells whose dropoff frequency is above a minimum threshold. It is interesting to note that cluster 1 (blue cells) has grouped most of the important transportation-related destinations together (e.g., the airport, bus and train stations).

2.3.2. Splitting via Road Hierarchy. A natural way to cluster a digital map is by using a *hierarchical* road network. In the United States, roads are classified into 4 levels, whereas in the European Union, roads are classified into 13 levels. A popular approach first divides the road network into groups enclosed by the highest-level roads. Each of these areas is then further subdivided into groups enclosed by the second highest level roads, and so on. The result is a hierarchical partition of the road network that uses roads as group boundaries. This approach was used in Gonzalez et al. [2007], Yuan and Zheng [2010], Zheng et al. [2011], Liu et al. [2011], and Yuan et al. [2012a, 2012b].

Fig. 4. Clustering using dropoff frequency as TimeFeature (only cells with a frequency above threshold plotted). ©Google (2011).

In a similar vein, Li et al. [2009, 2011c] constructed a three-level road hierarchy by classifying roads according to the frequency of their use. Sanders and Schultes [2005] use graph-theoretic methods to construct a road hierarchy automatically. However, their approach is designed for large-scale digital maps that include various cities, as these cities will usually be far apart, facilitating the detection of highways between them. As such, their method will most likely not be suitable for single-city digital maps.

2.4. Dealing with Data Problems

In working with real GPS data, there are many problems that arise, which must be properly addressed to guarantee reliable results. We review some of these problems along with possible approaches to overcoming them. Although far from being an exhaustive list, we hope that it provides an insight into the type of problems of which researchers must be aware in order to reduce "tainted" results.

2.4.1. Missing Data Entries. Occasionally, the GPS data is not received properly, with a lapse of several minutes or even hours between entries; the reasons for this lapse could be errors in the GPS device or a lack of GPS signal due to location (e.g., in an underground parking area). This results in a big physical jump with no information about the car's movement throughout this time. Depending on the problem being addressed, this data can either be left as it is, the trajectory can be split into two (or more) parts, or the trajectory can be truncated at the point before the jump occurred.

2.4.2. Erroneous Data Entries. Certain entries contain erroneous data, such as erroneous lat/lon or time entries. Most of the time, these are isolated entries that can be easily identified by using contextual information from the surrounding entries. A simple way to overcome these erroneous entries is by either adopting one of the methods mentioned in the last subsection or fixing the erroneous entry by extrapolating from the surrounding entries.

2.4.3. Multiple Drivers. A taxi may be driven by more than one driver, but this information cannot be obtained from the raw GPS traces. As we will see next, we are often interested in ranking the performance of drivers. In order to do so properly, we must be able to detect which driver is currently active. This is a fairly complex problem, and there are no existing solutions. A possible approach is to first detect the area where the drivers switch by monitoring areas that are always visited at the same time. However, since driver information is not available, we have no way of verifying the success rate of such a method.

2.4.4. Occupied Flag Improperly Set/Detected. The state of the occupied flag may not be properly set, either because the taxi driver does not set his indicator properly or

because there is a fault in the detection device. This may result in certain taxis being continuously occupied or vacant. When attempting to extract information from occupied/vacant trips, this type of problem can negatively impact the obtained statistics. One possibility for dealing with this problem is to compute the proportion of time the taxi is occupied/vacant and exclude any taxis that have extreme values (e.g., being occupied more than 95% of the time).

A related, but more difficult, problem is when a taxi picks up a new passenger immediately after dropping off another. Because of the rate at which GPS entries are received, we may not be able to determine when one trip ended and when the other began. This can be observed in popular transport areas such as the airport: a taxi dropping off a passenger at the airport may find a new passenger immediately.

2.4.5. "Sleeping" Taxis. Although having multiple drivers allows some taxis to operate at all hours of the day, single-driver taxis must stop to sleep at certain points. It is important to distinguish between a sleeping taxi and a taxi that is waiting for a passenger. In a similar manner to what was proposed previously, one could begin by detecting areas where the taxi is always parked at the same time.

3. SOCIAL DYNAMICS

Having finished surveying the work in preparing the taxi data, we can begin the survey of existing data mining research. We begin with social dynamics, which usually refers to the study of the collective behavior of a group of individuals. In the context of this work, it refers to the study of the collective behavior of the taxi drivers and passengers of a city, based on their observed movement in the city. This encompasses problems such as identifying hotspots, mobility patterns, and connectivity between regions. The problems discussed in this section are by no means an exhaustive list of what social dynamics entails. Indeed, as we will discuss in the Conclusion and Further Work section, one of the most important issues in this field is the proper definition of new problems and potential solutions.

Taxis provide a unique window into the dynamics of human movement in a city. As opposed to public transportation, which follows a predefined route, taxis take passengers between any two points. One can use the location and times where passengers were picked up and dropped off, as well as the trajectories followed, to analyze the activity levels throughout the city and the way people move around the city.

Before reviewing the work specific to taxi data, it is worth mentioning that other type of digital footprints, such as mobile phone data, have been used to study the human behaviors and social dynamics. This set of work uses a combination of various signals such as location traces (e.g., GSM and GPS), proximity information (Bluetooth), and communication logs (phone calls and SMS) to uncover human networks [Aharony et al. 2011], predict daily routines [Altshuler et al. 2012], investigate the relationship between human networks and personal decision making [Aharony et al. 2011], study long-distance travel patterns [Bekhor et al. 2011], analyze crowd mobility during special events [Calabrese et al. 2010a], investigate the occurrence of anomalous events [Candia et al. 2008], and investigate the predictability of human behaviour [González et al. 2008; Song et al. 2010a, 2010b]. The Reality Mining project [Eagle and Pentland 2006] is one of the representative studies and has generated a dataset of 100 users' locations and communication logs collected over a year, used by to discover daily routines [Farrahi and Gatica-Perez 2008, 2011] and friendship networks [Eagle and Pentland 2009]. Other well-known studies are MIT SENSEable City Lab's Mobile Landscapes [Ratti et al. 2009] and Eigenplaces [Calabrese et al. 2010b], which are examples of urban analysis for identifying hotspots and clustering similar areas in metropolitan areas based on cellular data. The Mobile Millenium project (http://traffic.berkeley.edu)

uses GPS data from cellular phones to provide real-time traffic information back to the mobile phone users.

We begin by discussing methods to extract hotspots in a city and characterize them based on the passenger pickup-dropoff activity.

3.1. Extracting Hotspots

The ability to detect the most frequented locations in a city can be useful for urban planning, public transportation route design, tourism and agencies, security agencies, amongst others. Taxis provide a precise indication of people's desired destinations. In this section, we discuss some methods for detecting a city's hotspots.

There have been extensive studies on using GPS trajectories from personal devices (e.g., cell phones) to detect significant places. Ashbrook and Starner [2003] define significant places as areas with a faded GPS signal (e.g., as would occur when inside a building) for a minimum amount of time. Alvares et al. [2007] define stops as polygons in $\mathbb{R}^2$ where trajectories spend a minimum amount of time. The authors assume that a set of candidate stops are given and present a simple algorithm for determining which are true stops. Palma et al. [2008] extend this by clustering sections of trajectories based on their speed, possibly suggesting more significant places than originally provided. Zheng et al. [2009] define stay points as areas (bounded by a distance threshold) where a user has stayed for a minimum amount of time. Aside from using mobile data, there is some related work using traces from other geopositioned devices for extracting interesting places. A representative study is the GeoLife project [Zheng et al. 2010], which collected the raw GPS trajectories (over various transportation modes) of 167 users over 3 years. This dataset was used by a number of projects for inferring transportation modes [Zheng et al. 2008], and for location and travel route recommendations [Zheng et al. 2009]. Calabrese et al. [2010b] use eigendecomposition on the time series frequencies of WiFi usage around MIT's campus and cluster regions around the campus based on their functionality. Girardin et al. [2008] discover the presence and movement of tourists using georeferenced photos and mobile phone data. Another well-known work is Dartmouth's CenceMe project [Miluzzo et al. 2008], which aims to create intelligent mobile sensor networks capable of sensing nearby friends and their current activity.

The works of greatest relevance to this article are those that use the traces from GPS-equipped vehicles. Wang et al. [2009] use passenger pickup and dropoff points to analyze the location and travel patterns to and from hotspots. Liu et al. [2010b] use vehicular speed information to quantify the "crowdedness" of an area, and define hotspots based on these crowdedness values. Yuan and Zheng [2010] define landmarks as the road segments most frequently traversed by taxi drivers; the purpose of these landmarks, however, is mainly to facilitate path planning. Chang et al. [2010] proceed by first filtering trajectories using contextual information (weather, etc.), then clustering GPS points into areas and finally defining a hotness score for each area according to the number of taxi requests divided by the size of the area. Li et al. [2011a] define hotspots as areas where there is a high level of passenger pickups and propose a method for predicting the amount of pickups at each hotspot by using a variant of the Auto-Regressive Integrated Moving Average (ARIMA), a well-known prediction method in time series analysis [Box et al. 2008].

With taxi datasets, we know with reasonable accuracy where passengers have been dropped off. This information is very useful, as it can be used to directly detect places of interest. By simply counting the number of dropoffs at different areas from Ψ, we can directly compare the importance of different locations. We can add further contextual information such as time of day, season of the year, and so forth, to uncover more meaningful results. In Figure 5, we display the weekday hotspots using the digital map decomposition. It is somewhat surprising that the airport (at the bottom right of

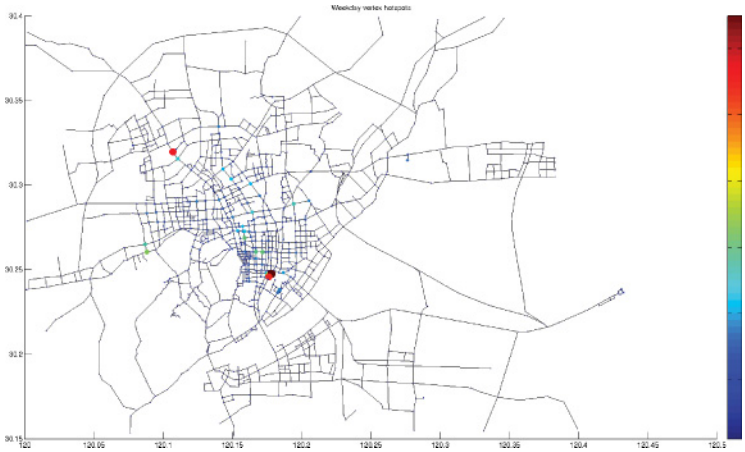

Fig. 5. Weekday vertex hotspots obtained by number of dropoffs in (or close to) each vertex. Circle sizes are proportional with the number of dropoffs.

the map) does not have a higher rating. This is most likely because when passengers are dropped off at the airport, taxi drivers immediately pick up a new one. Since there may not be any unoccupied entries between these two passengers, it will appear that the original passenger was never dropped off at the airport. This was discussed in Section 2.4.

3.2. Urban Computing

Urban computing is concerned with understanding the mobility patterns of a city's population. This type of research can be done in many different ways, including characterizing locations based on their functions, measuring the frequency of different trajectories between two points, and measuring the "linkage strength" between two areas in a city. The results uncovered are useful for many purposes, including determining the effectiveness of the road network and public transportation systems, and providing guidance for changes or additions that need to be made to the road network and/or public transport system in order to accommodate the population's movement patterns.

In order to ensure that roads are capable of satisfying traffic demand, researchers have sought to extract frequent trajectories used by drivers. Girardin et al. [2008] uncover the movement and presence of tourists using georeferenced photos and mobile phone data. Wang et al. [2009] use passenger pickup and dropoff points to analyze the travel patterns to and from hotspots. Chen et al. [2010a] use taxi trajectories to discover frequent trajectories between a given source and destination. Zhang et al. [2012] find patterns in origin-destination flows in a large city in China; the authors cluster together similar traces and find that traces in the same cluster have origins and destinations with similar social functions. Liu et al. [2009b] use taxi traces and smart cards from the bus and metro to spatially and temporally quantify, visualize, and examine urban mobility patterns; for taxi data, they present visualizations of the volume of taxi trips throughout the week, the strength of connections between different areas of city, and the change in taxi usage between the weekend and weekday.

Zheng et al. [2011] discover inefficient connectivity between two regions by looking at actual versus expected distance required to travel between these two regions, as well as the expected speed and actual volume of traffic, to determine whether their level of connectivity satisfies the demand of travel between them. They evaluate their

results using a taxi dataset in Beijing and demonstrate that the flawed areas uncovered by their algorithm agree with a new subway line added in the same area at a later date. Veloso et al. [2011b] present a visualization of the spatiotemporal variation, main pickup and dropoff areas, and busiest periods of taxi operation in Lisbon, Portugal; the same group also argued that trip distance, duration, and income follow Gamma and Exponential distributions [Veloso et al. 2011c]. Liu et al. [2012c] extract the temporal variations in taxi pickup and dropoff patterns and demonstrate that they correlate well with the "land use" in those areas (i.e., commercial, residential, recreational). Giannotti et al. [2011] present a platform for mobility analysis of GPS traces (M-Atlas) that can be used to query and plot the activity levels throughout the week, trip length and travel time distributions, clustering of routes to/from a particular point, and linkage between different areas. The authors also provide a method for detecting special events (such as sport matches), along with the trips related to this event. As an extension to using solely taxi GPS data, Hu et al. [2009] propose combining taxi traces and mobile phone positioning systems to model population density and travel time, distance, and frequency.

Taking this idea further, Yuan et al. [2012a] combine a hierarchical road network decomposition, general human mobility (gathered from vehicle, mobile, or social network traces) and points of interest of each region (i.e., restaurants, shopping malls) to uncover the functionality of different regions. Their approach uses a topic model-based method to identify the different functions by treating a region as a document, a region's function as a topic, human mobility amongst regions as words, and a region's points of interests as metadata. The model used is a generative one, and they proceed to cluster different areas according to their "topic" distribution and quantify the "intensity" of a region's function using Kernel Density Estimation. The authors evaluate their algorithm using a large dataset over 2 years in Beijing consisting of a points of interest database and a set of trajectories generated by 12,000 taxis; in comparison with other approaches, their proposed method produces very good results.

There has also been some work in characterizing the physical laws of human movement by means of taxi trajectories. This type of work has its roots in biology where the movement of animals are studied. It has been observed that the movement of many animals follow a Lévy flight model, which is a random walk that generalizes Brownian motion. It can be detected by verifying whether the jump length follows a power law behaviour. Chen et al. [2010b] studied the distribution of travel time and distance of taxi trips and showed that they can be approximated by a power law distribution; additionally, they also showed that most trips are short in both time and distance. However, Jiang et al. [2009] had previously shown that using taxi data in order to provide evidence of human mobility as a Lévy flight is mainly due to the underlying street network. Liu et al. [2012b] study this problem on a large 7-day database of taxi GPS traces in Shanghai and argue that trip distances do follow the power law distribution, but that the direction distribution is not uniform.

4. TRAFFIC DYNAMICS

The collective movement of vehicles throughout a city's network results in congestion levels that vary throughout different areas and time periods. These congestion levels influence important factors for drivers such as the travel time between two points, the expected speed, and potential adverse traffic events (e.g., accidents). Although we discuss some of these measures in this section, they are by no means exhaustive but are merely a reflection of the work that has been published so far. An understanding of the dynamics of these congestion levels, or the traffic dynamics, can be useful for the type of research discussed in the previous section, as well as for operational dynamics, which we will discuss in Section 5. Additionally, they can be used to analyze certain side

effects of vehicle use, such as estimating pollution levels in a city [Gühnemann et al. 2004]. Given that taxi drivers are continuously driving around the city, the collected GPS traces are a natural source for estimating the travel time between two points. Zhang et al. [2007] demonstrated that using taxi GPS data for estimating travel time and speed conditions is practical by performing an error analysis of taking simple averages over historical data.

Monitoring and predicting traffic conditions can provide indications of the activity levels in a city and can be useful for streamlining the flow of vehicles to reduce congestion levels. It has been observed that traffic generally follows a regular pattern throughout the day, and many groups have studied these patterns to obtain a better understanding of these dynamics. Researchers have used a vast array of different methods for estimating traffic conditions.

Wen et al. [2008] use GPS-equipped taxis to analyze traffic congestion changes around the Olympic games in Beijing; note that this is an ex post facto analysis of traffic conditions. Schäfer et al. [2002] used GPS-enabled vehicles to obtain real-time traffic information. By considering congested roads as those where the velocity is below 10km/hr, the authors demonstrate that a visualization of traffic conditions around the city can be used to detect congested and blocked road segments. Giannotti et al. [2011] detect traffic jams by searching for groups of cars close together that are all moving slowly. Gühnemann et al. [2004] use GPS data to construct travel time and speed estimates for each road segment, which are in turn used to estimate emission levels in different parts of the city. Their estimates are obtained by simply averaging over the most recent GPS entries. Šingliar and Hauskrecht [2007] studied two models for traffic density estimation: conditional autoregressive models and mixture of Gaussian trees. This work was designed to work with a set of traffic sensors placed around the city and not with GPS-equipped vehicles. Peng et al. [2012] decompose a city's traffic flow into a linear combination of three types of trips: travel between residential and work areas, travel between work areas, and leisure trips; they propose a method for finding these three coefficients, thereby producing a rough estimate of traffic flow. Lippi et al. [2010] use Markov logic networks to perform relational learning for traffic forecasting on multiple simultaneous locations, and at different steps in the future. This work is also designed for dealing with a set of traffic sensors around the city. Su and Yu [2007] used a Genetic Algorithm to select the parameters of a SVM, trained to predict short-term traffic conditions. Herring et al. [2010] use Coupled Hidden Markov Models [Brand 1997] for estimating traffic conditions on arterial roads. They propose a sophisticated model based on traffic theory that yields good results. Yuan et al. [2011a] used both historical patterns and real-time sensory information to predict traffic conditions. However, the predictions they provide are between a set of "landmarks" that is smaller than the size of the road network. Pang et al. [2011] propose using an adaptation of likelihood ratio tests (a technique mainly used in epidemiological studies) to describe traffic patterns and uncover unexpected events. Castro et al. [2012] propose a method to construct a model of traffic density and automatically determine the capacity of each road segment using a large database of taxi GPS traces; by pairing these two pieces of information, one can obtain accurate predictions of future traffic conditions and potential traffic jams.

Liu et al. [2011] aim to uncover causal relationships between regions considered as traffic outliers. The authors first decompose a city into a set of linked regions, where the links are obtained from taxi trajectories, and identify as outliers those links that are most different from both their spatial neighbours (using standard Euclidean distance) and their temporal neighbours (using Mahalanobis distance). Using this information, the authors present an algorithm for constructing outlier trees, thereby

revealing causal relationships amongst outliers. They further propose another algorithm that generates the most frequent subtree from the outlier trees, which can potentially reveal underlying problems in the existing road network. In their experiments, the authors' proposed method was successfully able to detect two known events that produced anomalous traffic patterns.

Accurate estimates of the travel time between two points in a city can be used for many different purposes, such as fare estimation and path planning. Blandin et al. [2009] use kernel methods [Scholkopf and Smola 2002] to obtain a nonlinear estimate of travel times on "arterial" roads; the performance of this estimate is then improved through kernel regression. Lou et al. [2009] use taxi trajectories to estimate turn probabilities at different intersections and the average speed of road segments at different times. Yuan and Zheng [2010] propose constructing a graph whose nodes are *landmarks*. Landmarks are defined as road segments frequently traversed by taxis. They propose a method to adaptively split a day into different time segments, based on the variance and entropy of the travel times between landmarks. This results in an estimate of the distributions of the travel times between landmarks. Balan et al. [2011] use simple arithmetical means of historical trajectories to estimate the travel distance between two points. They use both grid decomposition and a dynamic decomposition method based on the k-nearest neighbour method to group together similar start/end locations.

5. OPERATIONAL DYNAMICS

The information extracted from taxi GPS traces can also be used for the benefit of drivers. In this section, we discuss some existing methods for providing useful information to drivers and suggest some future avenues of research.

5.1. Ranking Drivers

Given the large number of taxis available, it is inevitable that some will be "better" than others. It follows that in order to extract useful information from these drivers, we should first rank their performance. The ranking criteria used depends on the type of behavior being analyzed. In Section 5.2, we discuss methods to uncover efficient strategies for finding new passengers. There are a number of methods for ranking drivers based on their efficiency of finding passengers, and we discuss a few next.

—One approach ranks drivers based on their average daily income, since one could assume that drivers generating high levels of profit are efficient at finding high-revenue passengers quickly. This should, of course, be normalized with respect to the operating time of each taxi, as many taxis are not operational 24 hours per day. This approach was used in Liu et al. [2010a] and Yuan et al. [2011a, 2011b]. Liu et al. [2009a] performed an analysis and comparison of top and ordinary drivers, ranked according to daily income, with the hope of revealing the top drivers' mobility intelligence.
—Li et al. [2011b] rank drivers based on the occupied distance covered by each taxi, as this is an indication that they know where to pick up "good" passengers that result in longer trips.
—An alternate method ranks drivers based on the proportion of time they spent occupied; that is, if the taxi is occupied most of the time, this is a good indication that the driver is effective at finding new passengers [Ge et al. 2010].
—One can also rank drivers based on the average time that it takes to find a new passenger, as this is a direct indication of the driver's passenger-finding efficiency.

In Section 5.3, we discuss methods to use the information in our dataset for path planning between two points. As discovered by Liu et al. [2010a], top drivers usually

take their passengers along the fastest route to their destination in order to get a new passenger quickly. Thus, ranking drivers based on their average daily income (as mentioned earlier) would be an appropriate ranking for this problem. An alternate method of ranking would be to rank the drivers according to the average time taken between one or more common source-destination pairs. This method has the disadvantage that the reliability of the ranking depends on how frequently passengers are driven between the chosen source-destination pairs.

5.2. Passenger/Taxi-Finding Strategies

Discovering good strategies for drivers to find new passengers or for passengers to find a taxi is a problem that has previously been investigated by a number of groups. Most papers have focused on finding demand hotspots to direct the navigation of unoccupied drivers (or waiting passengers). Chang et al. [2008] find demand hotspots by extracting the time and environmental contexts of a set of taxi requests, clustering these requests using k-means and agglomerative hierarchical clustering, and finally, ranking these clusters. Palma et al. [2008] use the speed of vehicles in a dataset of trajectories to find "interesting places" by means of a density-based clustering algorithm. Yue et al. [2009] use simple nearest-neighbor clustering to group taxi pickup and dropoff points and discover attractive areas as well as the attractiveness amongst different areas. Veloso et al. [2011a] explore the relationships between dropoff and pickup locations. Phithakkitnukoon et al. [2010] use a grid decomposition and a naive Bayesian classifier to predict vacant taxis in different areas. In addition to aiding passengers seeking taxis, this can be used to help guide drivers to areas with high demand but few vacant taxis. Ge et al. [2010] cluster the pickup points of the top drivers to use as recommended pickup points for other drivers. Hu et al. [2012b] extend this idea by creating a pickup tree with the pickup points with highest probability; the authors argue that this method is more suitable for situations where there is a set of vacant taxis (as opposed to a single one) in the same area. Zheng et al. [2012] model the probability of taxis leaving their current road segment as a Nonhomogeneous Poisson Process [Ross 2006] and use this model to estimate the waiting time for taxis at different locations and at different times; these estimates are then used to provide a recommender system for people searching for taxis.

Liu et al. [2010a] use k-means clustering to uncover the spatiotemporal preferences of the top drivers in a city. Statistical analysis reveals that most top taxi drivers (ranked according to profit) choose similar spatiotemporal areas. The authors discovered the somewhat surprising facts that top drivers strived to drop off passengers as quickly as possible in order to serve as many passengers as possible; additionally, they chose to operate in areas *other* than the Central Business District.

Takayama et al. [2011] use source/destination data of occupied taxis over a 2-year period and a series of survey results from the drivers to propose promising "waiting/cruising" locations to taxi drivers. Their method is based on surveys given to drivers, which is a very inefficient way of obtaining data, prone to human error, and difficult to continue indefinitely. Lee et al. [2008] use k-means clustering to split a road network into different areas. They then perform a temporal analysis to create a time-dependent pickup pattern within each area. Their analysis suggests that taxis should go to the nearest area with demand to pick up new customers. The simple approach is able to find clusters with highest demand. However, as Liu et al. [2010a] demonstrated, in order to maximize profit, a taxi driver may not necessarily want to base his choice solely on demand. A balance between profit maximization and demand coverage is necessary.

Li et al. [2011b] categorize the observed passenger-finding strategies based on time, location, whether they are "hunting" or "waiting", and whether the driver remains in a local area or travels a longer distance to find a new passenger. The authors then use

a form of Support Vector Machine (SVM), L1-Norm SVM [Bi et al. 2003], to determine, based on the current time and location, whether the driver should hunt, wait, stay locally, or travel a longer distance.

Yuan et al. [2011a, 2011b] automatically extract "waiting areas" for taxis based on the distance between consecutive GPS points. The authors then compute the probability of picking up a passenger based on the current time and the road segment or waiting area. This information is used to provide a recommendation system for drivers and passengers. Similarly, Li et al. [2011a] use extracted hotspots and distance to hotspots to help drivers find new passengers. Qian et al. [2012] extract pickup points and formulate the taxi-routing problem as a Markov Decision Process [Puterman 1994] between pickup points; however, this approach ignores the destination desired by each new potential customer and is evaluated only on success probabilities extracted from prior data.

As mentioned previously, most of these works focus on finding areas where taxis can wait for new passengers. However, as shown in Takayama et al. [2011] and Li et al. [2011b], a driver may often want to hunt or cruise. Powell et al. [2011] construct a spatiotemporal profitability map based on historical data to guide taxis on a local basis. Yamamoto et al. [2010] provide routing strategies for multiple taxis using fuzzy clustering mechanisms. Hu et al. [2012a] formulate the taxi driver's task of hunting for new passengers as a decision problem at each intersection and propose solving it using probabilistic dynamic programming. Nevertheless, it is not clear whether a clear "microstrategy" for finding passengers can be extracted by mining past taxi trajectories: the strategies employed by top drivers would have to be fairly consistent or *predictable*. Veloso et al. [2011b] perform a predictability analysis of the next pickup area given dropoff features. Their results show there is only a 54% predictability rate, suggesting that hunting/cruising trips are largely random.

A new research direction is taxi ridesharing, which has been explored by a number of groups. Tao [2007] give an overview of one such service, algorithms for rideshare matching, and an empirical evaluation of a field trial in Taipei. Chen [2010] propose a dynamic ride-sharing system for Vehicular Ad-hoc Networks (VANETs) and conduct a simulation to estimate the fuel savings resulting from such a system. d'Orey [2012] also perform an empirical evaluation of ridesharing using simulations. Lin et al. [2012] propose an algorithm for optimizing routing of a rideshare service that aims to minimize operating costs while maximizing customer satisfaction. Finally, Ma et al. [2013] analyze the potential passenger coverage increase and travel mileage decrease that could result from offering a taxi-ridesharing service. They further propose a dynamic ridesharing service that achieves its goal by fast taxi searching and lazy shortest path algorithms.

5.3. Route Planning

The travel time and traffic estimates discussed in Section 4 can be used directly to perform path planning from a source to a destination location. The generalized routing problem has been studied extensively for (at least) four decades. Popular techniques that have often been used are dynamic programming [Cooke and Halsey 1966], variations of Djikstra's algorithm [Ding et al. 2008], and variations of the A^* algorithm [Kanoulas et al. 2006].

Edelkamp and Schrödl [2003] construct a map and use existing methods (such as Djikstra's algorithm) to perform planning. Šingliar and Hauskrecht [2008] propose several approximation strategies based on Monte Carlo sampling that can be used to solve route-planning problems in stochastic transportation networks, formalized as semi-Markov Decision Processes [Puterman 1994]. Li et al. [2009a, 2011c] construct a

hierarchy of roads based on frequency of use and perform planning from a source to a target by trying to travel through the highest hierarchy roads. Yuan et al. [2011a] combine historical traffic patterns, real-time traffic information, and driver behaviours to compute shortest-time driving routes.

There has been some work on finding routes within a prespecified time frame. Kanoulas et al. [2006] extends A^* for finding shortest paths within a specified interval (rather than a fixed time). Using stochastic modeling of travel delay on road networks, Lim et al. 2010 present a stochastic motion-planning algorithm that can be used to find paths that maximize the probability of reaching a destination within a particular travel deadline.

Since top drivers usually know the city's road network very well, by observing their behavior we may be able to uncover good path-planning strategies that go beyond the techniques mentioned previously. Gonzalez et al. [2007] find the fastest path between two points by partitioning the map into a road hierarchy, extracting frequently traveled road segments to use as "hints" and precomputing high-benefit paths for each area. The data is mined from road speed sensors. Yuan and Zheng [2010] and Yuan et al. [2013] construct a time-dependent landmark graph based on a large dataset of taxi trajectories. The routing algorithm first finds a rough route on the landmark graph, then this is refined to a route on the underlying road network. By estimating travel time distributions, the authors allow travel times to behave stochastically, which may yield more accurate representations.

Bastani et al. [2011] propose defining new transportation routes by mining through and combining multiple taxi trajectories. The authors suggest that these new routes could be used by a mini-shuttle transportation system that lies somewhere between taxis and buses.

Chen et al. [2013] leverage taxi GPS traces to suggest nightly bus routes. Their approach works by first clustering dense passenger pickup/dropoff areas into hotspots and then using these hotspots as candidate bus stops. The authors propose two heuristic algorithms to generate candidate bus routes and evaluate their effectiveness on a large taxi GPS dataset.

5.4. Anomaly Detection

Given a fixed source and destination, there are undoubtedly certain trajectories favoured by most taxi drivers. By collecting the trajectories from many taxis over an extended period of time, we may be able to automatically recognize not only these "normal" trajectories but also "anomalous" trajectories. An anomalous trajectory can be caused by external factors such as accidents or the closure of a main road, but it may also be caused by fraudulent drivers trying to charge more money from passengers. The ability to automatically detect anomalous trajectories can thus enable one to automatically detect adverse traffic events, as well as preventing drivers to take advantage of passengers unfamiliar with the city.

Li et al. [2009b] identify outlier road segments by detecting drastic changes between current data and historical trends. Liao et al. [2010] use conditional random fields to label anomalous taxis, coupled with an active learning scenario, where human interaction can help guide the learning. Balan et al. [2011] label trajectories as anomalous by a simple mechanism: any trajectory with a distance twice as long as the straight line distance between the start and end positions, or any trajectory with an average speed lower than 20 km/hr or higher than 100 km/hr.

Chawla et al. [2012] propose a two-step method for inferring the root cause of traffic anomalies. First, the city is divided into disjoint regions, and anomalous links between different regions are identified by detecting deviations from historical norms. In the

second step, the authors combine a generative model of the traffic flow in the city and link adjacency matrices to uncover the probable sources of the observed anomalies.

Zhang et al. [2011b] propose *iBAT*, a method based on isolation trees and a grid decomposition, to solve this problem. The authors maintain a set of historical trajectories and determine whether new trajectories are *isolated* from this set by randomly selecting grid cells from the new trajectory and determining how many of the historical trajectories also contain this grid cell. Since the method is based on sampling, the process must be repeated a number of times for each trajectory in order to obtain an *anomaly score* that indicates the degree of anomalousness of the new trajectory. Through the use of a testing set of manually labelled trajectories, the authors verified the accuracy of their proposed method.

Despite its accuracy, *iBAT* has a number of shortcomings. The most important one is that it only works with completed trajectories, disqualifying it from being used for real-time fraud detection. Additionally, because it is selecting grid cells independently, it does not maintain contextual information and is thereby unable to detect *looping* trajectories—that is, trajectories that go back on its steps and return along the same route. To address these issues, Chen et al. [2011] propose *iBOAT*, an algorithm that can detect anomalous trajectories in real time and avoid the shortcomings of *iBAT*. The main idea behind *iBOAT* is to compare subsequences of the new trajectory against subsequences of the historical trajectories. If there is enough *support*, they increase the size of the subtrajectory from the trajectory we are testing; otherwise, the point is labelled as anomalous and the process is repeated from the next point. By doing so, contextual information is preserved without sacrificing accuracy, as confirmed by the empirical evaluation. Sun et al. [2012] extend this work by proposing the use of an inverted indexing mechanism to guarantee real-time monitoring of anomalous trajectories; the authors further perform an analysis of the types of anomalous trajectories observed in the dataset.

More recently, Ge et al. [2011] proposed a similar method for detecting taxi fraud. Their method uses a grid decomposition and complete trajectories in a similar way as done in Zhang et al. [2011b] and Chen et al. [2011]. They compute two pieces of evidence for detecting anomalous trajectories. The first involves computing the independent components (using Independent Component Analysis) of a set of trajectories and then computing the *coding cost* (which is essentially the entropy) of a trajectory's independent components. The second is a method for determining the expected distance for the most common routes and computing how much a trajectory's distance differs from the norm. These two pieces of evidence are combined using Dempster-Schafer theory. Although their experimental results fail to convince the reader that their method provides an advantage over standard density-based methods, they provide mechanisms for differentiating between malicious detours and detours due to traffic interruptions or poor knowledge of the area.

5.5. Route Prediction

GPS devices are mainly used for providing driving directions after specifying a desired destination. There has been some recent work on using GPS trajectories to *predict* a user's route and/or destination based on historical information. The ability to predict routes can be useful for a number of problems, including automatically providing driving directions without user input and providing warnings about possible hazards along the predicted path. Lou et al. [2009] compute a simple distribution over possible destinations given the taxi's source. Patterson et al. [2003] learn a Bayesian model of a traveller's current mode of transportation and most likely route. Krumm and Horvitz [2006] also use Bayesian inference to predict the drivers destination as the trip progresses based on both the drivers history of destinations and the trips for a group

of drivers. Liao et al. [2007] use a hierarchical Markov model to predict a user's daily movements, as well as automatically detect significant locations. Froehlich and Krumm [2008] exploit the regularity of regular drivers to predict a driver's end-to-end route, using the driver's history of trajectories. Ziebart et al. [2008] use inverse Reinforcement Learning [Ng and Russell 2000] to predict turns, routes, and the destinations of a driver. Monreale et al. [2009] construct a decision tree from *trajectory patterns* [Giannotti et al. 2007] to predict the next location of moving objects; however, they use the information from *all* previous trajectories through a particular area to form their predictions.

More recently, Xue et al. [2013] proposed an approach for overcoming data sparsity issues when predicting destinations. Their approach decomposes historical trajectories into subtrajectories and then connects these to produce "synthesized" trajectories, allowing them to give predictions for an exponentially larger number of trajectories than possible when using only complete historical trajectories.

6. DEPLOYED SYSTEMS

Many of the ideas discussed so far can be exploited in the development of a commercial product. However, this has not been greatly pursued. In this section, we wish to list a few of the existing deployed systems that are used to manage GPS-enabled vehicles. Schäfer et al. [2002] use the real-time traffic information obtained from GPS-enabled vehicles to provide a realistic travel time and optimal route suggestion for individual and commercial users in Berlin, Vienna, and Nuremberg. Liao [2003] describes a taxi-dispatching system used in Singapore that uses the GPS location of taxis and the locations of nearby passengers. Balan et al. [2011] deployed their system to predict fares and trip durations in Singapore. CabSense (http://www.cabsense.com) is a smartphone app that uses the real-time information from GPS-equipped taxis to aid the passenger in finding a taxi. Cabspotting (http://stamen.com/clients/cabspotting) traces taxis in the San Francisco Bay area, which has been used by artists and researchers to visualize the underlying social dynamics of the city. TaxiTrack^{TM} is a real-time taxi tracking software that can facilitate taxi dispatching. Waze (http://www.waze.com) is a community-based GPS traffic and navigation app.

7. HISTORICAL PERSPECTIVE

We grouped the existing work into three main categories, and we would like to complement this with a historical overview of these works. Although GPS devices have been around since around the 1970s, for the first 30 years or so they were used almost exclusively for military purposes, since the signal for civilian use was purposefully degraded (known as Selective Availability). This degradation was removed in 2000, enabling their widespread use in civil sectors. It should be noted that the limitations imposed by Selective Availability had been overcome a few years earlier by the use of Differential GPS; however, their widespread commercial success followed the removal of Selective Availability. Since then, the use of GPS devices has risen dramatically, and they are present in a significant number of vehicles and mobile phones. In this section, we will only discuss works related to GPS-equipped taxis, as this is the focus of this article.

Although research papers using GPS-equipped vehicles began appearing around 2000, papers using GPS-equipped taxis only began appearing in 2003–2004, where they were used for a real-time dispatching system in Singapore [Liao 2003] and for real-time monitoring of traffic emissions [Gühnemann et al. 2004]. In 2007, Zhang et al. demonstrated that GPS-equipped taxis can be effectively used for estimating speed and travel time. In 2008, there were a couple of papers on uncovering hotspots for finding passengers [Chang et al. 2008; Lee et al. 2008], predicting the turns, routes, and

destinations of a particular user [Ziebart et al. 2008], and analyzing traffic conditions during the Olympic games in Beijing [Wen et al. 2008].

In 2009, there was a big surge in the number of papers using GPS-equipped taxis for a number of purposes, in particular for urban computing [Hu et al. 2009; Jiang et al. 2009; Liu et al. 2009b; Wang et al. 2009]. In addition, there was some work on detecting outlier road segments [Li et al. 2009] and automatic map construction [Lou et al. 2009].

In 2010, Microsoft Research Asia addressed the problems of finding hotspots and estimating travel time in order to provide driving directions [Yuan and Zheng 2010]; Chang et al. [2010] and Liu et al. [2010b] also proposed mechanisms for finding hotspots, whereas Herring et al. [2010] came out with a method for modeling traffic conditions on arterial roads. Researchers from MIT performed an analysis of passenger-finding strategies [Liu et al. 2010a] and methods for predicting vacant taxis in a city [Phithakkitnukoon et al. 2010]; Yamamoto et al. [2010] also provided routing strategies for taxis searching for new passengers. Finally, researchers from Zhejiang University produced a couple of papers on urban computing [Chen et al. 2010a, 2010b].

In 2011, there was an increased number of papers focusing on this area of research, especially in regards to passenger-finding strategies [Li et al. 2011a, 2011b; Powell et al. 2011; Takayama et al. 2011; Veloso et al. 2011a, 2011b; Yuan et al. 2011b; Yuan et al. 2011c] and urban computing [Qi et al. 2011; Veloso et al. 2011b, 2011c; Zheng et al. 2011b]. Additionally, there was some work on route planning [Yuan et al. 2011; Bastani et al. 2011], traffic monitoring [Liu et al. 2011; Pang et al. 2011], anomaly detection [Zhang et al. 2011; Ge et al. 2011; Chen et al. 2011], and traffic dynamics [Balan et al. 2011].

We have seen a continuation of a large body of work throughout 2012 and 2013, with papers related to map construction [Yuan et al. 2012b; Biagioni and Eriksson 2012; Liu et al. 2012a], urban computing [Yuan et al. 2012a; Zhang et al. 2012; Liu et al. 2012; Liu et al. 2012], traffic dynamics [Peng et al. 2012; Castro et al. 2012], passenger-finding strategies [Hu et al. 2012a, 2012b; Qian et al. 2012; Zheng et al. 2012; Ma et al. 2013], route planning/prediction [Xue et al. 2013; Chen et al. 2013; Yuan et al. 2013], and anomaly detection [Chawla et al. 2012].

For ease of reference, we summarize the previous research work according to the time line in Tables II and III. In addition, we group them by categories in Tables IV, V, VI, VII, and VIII.

8. CONCLUSION AND FUTURE WORK

GPS-equipped taxis can be viewed as pervasive sensors, and the large-scale GPS traces produced allow us to reveal many valuable facts about the social and community dynamics. In this article, we performed an exhaustive survey of existing work and grouped them into three categories: social dynamics, traffic dynamics, and operational dynamics. We first provided a formal specification of the datasets and presented an overview of different mechanisms for preprocessing the data to decompose the city.

Grid decompositions are a popular decomposition used by many research groups, but there can be many improvements on this simple decomposition that can better adapt to the road network structure and dynamics. As a simple example, a more adaptive grid decomposition can be achieved if one uses the "shape" of the GPS entries to guide the size and placement of the various grids, providing more accuracy to areas with higher GPS entry frequency. This effect can be achieved by using data structures such as Quad Trees [Finkel and Bentley 1974], Binary Space Partitioning (BSP) Trees [Fuchs et al. 1980], or R-Trees [Guttman 1984] (which allows overlapping regions).

Social dynamics refers to the study of the collective behaviour of a city's population, based on their observed movement in the city. The research along this line includes

Table II. Timeline of Papers Discussed

Year	Category	Reference
2003	Deployed systems	[Liao 2003]
2004	Traffic dynamics	[Gühnemann et al. 2004]
2007	Traffic dynamics	[Zhang et al. 2007]
	Map construction	[Li et al. 2007]
	Passenger/taxi-finding strategies	[Tao 2007]
2008	Passenger/taxi-finding strategies	[Chang et al. 2008] [Lee et al. 2008]
	Traffic dynamics	[Wen et al. 2008]
	Route prediction	[Ziebart et al. 2008]
2009	Map construction	[Lou et al. 2009]
	Urban computing	[Hu et al. 2009] [Jiang et al. 2009] [Liu et al. 2009b] [Wang et al. 2009]
	Traffic dynamics	[Lou et al. 2009]
	Passenger/taxi-finding strategies	[Yue et al. 2009]
	Anomaly detection	[Li et al. 2009b]
	Route planning	[Li et al. 2009a]
	Route prediction	[Lou et al. 2009]
2010	Map construction	[Yuan et al. 2010]
	Urban computing	[Chen et al. 2010a] [Chen et al. 2010b]
	Finding hotspots	[Liu et al. 2010b]
	Traffic dynamics	[Herring et al. 2010]
	Passenger/taxi-finding strategies	[Liu et al. 2010a] [Phithakkitnukoon et al. 2010] [Yamamoto et al. 2010] [Ge et al. 2010] [Chen 2010]
	Route planning	[Yuan and Zheng 2010]
	Anomaly detection	[Liao et al. 2010]
2011	Urban computing	[Qi et al. 2011] [Veloso et al. 2011c] [Veloso et al. 2011b] [Zheng et al. 2011]
	Traffic dynamics	[Balan et al. 2011] [Liu et al. 2011] [Pang et al. 2011]
	Passenger/taxi-finding strategies	[Li et al. 2011b] [Li et al. 2011a] [Powell et al. 2011] [Takayama et al. 2011] [Veloso et al. 2011a]
	Route planning	[Yuan et al. 2011a] [Li et al. 2011c] [Bastani et al. 2011]
	Anomaly detection	[Zhang et al. 2011b] [Ge et al. 2011] [Chen et al. 2011]

Table III. Timeline of Papers Discussed

Year	Category	Reference
2012	Map construction	[Yuan et al. 2012b] [Biagioni and Eriksson 2012] [Liu et al. 2012a]
	Urban computing	[Yuan et al. 2012a] [Zhang et al. 2012] [Liu et al. 2012c] [Liu et al. 2012b]
	Traffic dynamics	[Peng et al. 2012] [Castro et al. 2012]
	Passenger/taxi-finding strategies	[Hu et al. 2012a] [Hu et al. 2012b] [Zheng et al. 2012] [d'Orey 2012] [Lin et al. 2012]
	Anomaly detection	[Chawla et al. 2012] [Sun et al. 2012]
2013	Passenger/taxi-finding strategies	[Ma et al. 2013]
	Route prediction	[Xue et al. 2013]
	Route planning	[Yuan et al. 2013] [Chen et al. 2013]

Table IV. Data Preparation Papers

Data preparation	
Data processing	[Alvares et al. 2007] [Wang et al. 2011]
Map construction	[Rogers et al. 1999] [Greenfeld 2002] [Edelkamp and Schrödl 2003] [Yin and Wolfson 2004] [Schroedl et al. 2004] [Brakatsoulas et al. 2005] [Worrall and Nebot 2007] [Chawathe 2007] [Li et al. 2007] [Cao and Krumm 2009] [Lou et al. 2009] [Fathi and Krumm 2010] [Chen and Krumm 2010] [Yuan et al. 2010] [Wang et al. 2011] [Yuan et al. 2012b] [Biagioni and Eriksson 2012] [Liu et al. 2012a]
Map decomposition	[Gonzalez et al. 2007] [Li et al. 2009a] [Yuan and Zheng 2010] [Li et al. 2011c] [Zheng et al. 2011]

Table V. Traffic Dynamics Papers

Traffic dynamics	
	[Schäfer et al. 2002]
	[Gühnemann et al. 2004]
	[Šingliar and Hauskrecht 2007]
	[Su and Yu 2007]
	[Zhang et al. 2007]
	[Wen et al. 2008]
	[Blandin et al. 2009]
	[Lou et al. 2009]
	[Herring et al. 2010]
	[Lippi et al. 2010]
	[Yuan and Zheng 2010]
	[Yuan et al. 2011a]
	[Balan et al. 2011]
	[Giannotti et al. 2011]
	[Liu et al. 2011]
	[Pang et al. 2011]
	[Peng et al. 2012]
	[Castro et al. 2012]

Table VI. Social Dynamics Papers

Social dynamics	
Extracting hotspots	[Ashbrook and Starner 2003]
	[Alvares et al. 2007]
	[Girardin et al. 2008]
	[Palma et al. 2008]
	[Calabrese et al. 2010b]
	[Zheng et al. 2009]
	[Wang et al. 2009]
	[Liu et al. 2010b]
	[Yuan and Zheng 2010]
	[Chang et al. 2010]
	[Li et al. 2011a]
Urban computing	[Girardin et al. 2008]
	[Jiang et al. 2009]
	[Froehlich et al. 2009]
	[Liu et al. 2009b]
	[Hu et al. 2009]
	[Wang et al. 2009]
	[Chen et al. 2010b]
	[Chen et al. 2010a]
	[Zheng et al. 2011]
	[Qi et al. 2011]
	[Veloso et al. 2011b]
	[Veloso et al. 2011c]
	[Giannotti et al. 2011]
	[Yuan et al. 2012a]
	[Zhang et al. 2012]
	[Liu et al. 2012c]
	[Liu et al. 2012b]

Table VII. Operational Dynamics Papers

Operational dynamics	
Passenger/taxi-finding strategies	[Tao 2007] [Chang et al. 2008] [Lee et al. 2008] [Yue et al. 2009] [Liu et al. 2010a] [Yamamoto et al. 2010] [Phithakkitnukoon et al. 2010] [Ge et al. 2010] [Chen 2010] [Li et al. 2011b] [Takayama et al. 2011] [Yuan et al. 2011b] [Yuan et al. 2011c] [Powell et al. 2011] [Li et al. 2011a] [Veloso et al. 2011b] [Veloso et al. 2011a] [Hu et al. 2012a] [Hu et al. 2012b] [Zheng et al. 2012] [d'Orey 2012] [Lin et al. 2012] [Ma et al. 2013]
Route planning	[Edelkamp and Schrödl 2003] [Kanoulas et al. 2006] [Gonzalez et al. 2007] [Ziebart et al. 2008] [Li et al. 2009a] [Lim et al. 2010] [Yuan and Zheng 2010] [Li et al. 2011c] [Yuan et al. 2011a] [Bastani et al. 2011] [Yuan et al. 2013] [Chen et al. 2013]
Anomaly detection	[Li et al. 2009b] [Liao et al. 2010] [Zhang et al. 2011b] [Ge et al. 2011] [Chen et al. 2011] [Chawla et al. 2012] [Sun et al. 2012]
Route prediction	[Patterson et al. 2003] [Krumm and Horvitz 2006] [Liao et al. 2007] [Giannotti et al. 2007] [Froehlich and Krumm 2008] [Ziebart et al. 2008] [Monreale et al. 2009] [Lou et al. 2009] [Xue et al. 2013]

Table VIII. Deployed Systems Papers and Web Sites

Deployed systems
[Schäfer et al. 2002]
[Liao 2003]
[Balan et al. 2011]
http://www.cabsense.com
http://stamen.com/clients/cabspotting
http://www.waze.com

hotspot extraction and urban computing. The approach chosen for extracting hotspots depends on their ultimate use (i.e., urban computing, route planning) and can be obtained by simply counting the pickup and dropoff events in an area or using more sophisticated clustering techniques. There are few papers using taxi-GPS data for urban computing. This is a very promising area of research that has many open avenues for exploration. As opposed to other problems where sophisticated algorithms are usually necessary, fairly simple algorithms are usually sufficient for urban computing. The challenge comes in properly defining questions and problems that one can answer by mining these trajectories. The methodologies used to answer these questions might be specific to this type of data, but the questions raised can be valuable on their own and may motivate research using other types of data.

Traffic dynamics studies the resulting flow of the vehicle movement through the road network. Monitoring and predicting traffic conditions can provide indications of the activity levels in a city. Accurate estimates of the congestion level in each road segment and the travel time between two points in a city are still open but valuable research topics.

Operational dynamics refers to the general study and analysis of taxi drivers' modus operandi. This new line of research focuses mainly on providing optimal route planning and passenger-finding strategies to taxi drivers by analyzing the tactics of good drivers. The performance of many of the algorithms presented in Section 5 directly depends on the quality of the ranking function. Of the rankings presented, the advantages and disadvantages of each are not clear. The algorithms we presented for anomaly detection are fairly simple, yet they yield fairly good results. We believe more sophisticated algorithms that take more factors into consideration (e.g., speed, weather conditions) can yield even superior performance.

Additionally, as mentioned previously, the ability to detect multiple drivers per taxi is an open problem that has not yet been addressed. We presented a simple approach for this problem in Section 2.4.3, but there are certainly more sophisticated methods to be discovered. The major difficulty with this problem is determining how to verify the accuracy of any algorithm proposed.

Taxi ridesharing [Tao 2007; Chen 2010; d'Orey 2012; Lin et al. 2012; Ma et al. 2013] is a promising new research direction that can have significant social and environmental impact. Given the rapid growth of cities and use of motorized vehicles, taxi ridesharing provides a promising mechanism to mitigate increasing road congestion.

Given that smartphones have become an essential element in the lives of most people in developed countries, it is clear that the usefulness of the research surveyed here will need to be integrated with the research in mobile data. Social networking services such as Facebook and Twitter have also become an essential part of people's lives that both reveals and influences people's behaviour. As such, an important avenue for future work will be integrating the research surveyed here with the research in social media. In particular, the work in social dynamics presented here has a strong relationship to some research on social networks such as Twitter [Demirbas et al. 2010; Culotta 2010;

Bollen et al. 2011], Facebook [Backstrom et al. 2010], Flickr [Crandall et al. 2010], and Brightkites [Li and Chen 2009].

The common underlying goal shared amongst the different works discussed is to obtain deep insights and useful applications to meet the real-world needs of end users, such as urban planners, taxi operators, taxi drivers, passengers, and regular city dwellers. Taxi GPS traces, despite being a very specialized type of digital trace, have already provided us with a rich dataset to uncover many hidden facts about the city, including social dynamics, traffic dynamics, and operational dynamics. Coupling this data source with other complementary data sources such as mobile phone logs, private and public transportation usage records, road network sensors, social media, and Internet applications will allow us to uncover many more interesting facets about the the city and its inhabitants. The ability to obtain these results in real time along with appropriate visualization interfaces can offer city planners and dwellers many insights and services that bring them closer to the vision of a smart city.

REFERENCES

Aharony, N., Pan, W., Ip, C., Khayal, I., and Pentland, A. 2011. Social fMRI: Investigating and shaping social mechanisms in the real world. *Perv. Mobile Comput. 7*, 6, 643–659.

Alt, H., Efrat, A., Rote, G., and Wenk, C. 2003. Matching planar maps. *J. Algorithms 49*, 262–283.

Altshuler, Y., Aharony, N., Pentland ("Sandy"), A., Fire, M., and Elovici, Y. 2012. Incremental learning with accuracy predictions of social and individual properties from mobile-phone data. In *Proceedings of the 1st International Workshop on Wide Spectrum Social Signal Processing*.

Alvares, L. O., Bogorny, V., Kuijpers, B., de Macedo, J. A. F., Moelans, B., and Vaisman, A. 2007. A model for enriching trajectories with semantic geographical information. In *Proceedings of the 15th Annual ACM International Symposium on Advances in Geographical Information Systems (GIS'07)*. 22:1–22:8.

Ashbrook, D. and Starner, T. 2003. Using GPS to learn significant locations and predict movement across multiple users. *Pers. Ubiq. Comput. 7*, 5, 275–286.

Backstrom, L., Sun, E., and Marlow, C. 2010. Find me if you can: Improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*. 61–70.

Balan, R. K., Khoa, N. X., and Jiang, L. 2011. Real-time trip information service for a large taxi fleet. In *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services (MobiSys'11)*. 99–112.

Bastani, F., Huan, Y., Xie, X., and Powell, J. 2011. A greener transportation mode: Flexible route discovery from GPS trajectory data. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS'11)*. 405–408.

Bekhor, S., Cohen, Y., and Solomon, C. 2011. Evaluating long-distance travel patterns in Israel by tracking cellular phone positions. *J. Adv. Transport. 47*, 4, 435–446.

Bi, J., Bennett, K., Embrechts, M., Breneman, C., and Song, M. 2003. Dimensionality reduction via sparse support vector machines. *J. Mach. Learn. Res. 3*, 1229–1243.

Biagioni, J. and Eriksson, J. 2012. Inferring road maps from GPS traces: Survey and comparative evaluation. *Transport. Res. Rec. 2291*, 61–71.

Blandin, S., Ghaoui, L. E., and Bayen, A. 2009. Kernel regression for travel time estimation via convex optimization. In *Proceedings of the 48th IEEE Conference on Decision and Control*. 4360–4365.

Bollen, J., Mao, H., and Pepe, A. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*. 450–453.

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. 2008. *Time Series Analysis*. Wiley.

Brakatsoulas, S., Pfoser, D., Salas, R., and Wenk, C. 2005. On map-matching vehicle tracking data. In *Proceedings of the 31st International Conference on Very Large Data Bases (VLDB'05)*. 853–864.

Brand, M. 1997. Coupled hidden Markov models for modeling interacting processes. Techrep. The Media Lab, Massachusetts Institute of Technology.

Calabrese, F., Pereira, F. C., Lorenzo, G. D., Liu, L., and Ratti, C. 2010a. The geography of taste: Analyzing cell-phone mobility and social events. In *Proceedings of the 8th International Conference on Pervasive Computing (Pervasive'10)*. 22–37.

Calabrese, F., Reades, J., and Ratti, C. 2010b. Eigenplaces: Segmenting space through digital signatures. *Pervasive Comput. 9*, 1, 78–84.

Candia, J., González, M. C., Wang, P., Schoenharl, T., Madey, G., and Barabási, A.-L. 2008. Uncovering individual and collective human dynamics from mobile phone records. *J. Phys. A-Math. Theor. 41*, 22.

Cao, L. and Krumm, J. 2009. From GPS traces to a routable road map. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS'09)*. 3–12.

Castro, P. S., Zhang, D., and Li, S. 2012. Urban traffic modelling and prediction using large scale taxi GPS traces. In *Proceedings of the 10th International Conference on Pervasive Computing (Pervasive'12)*. 57–72.

Chang, H.-W., Tai Y.-C., and Hsu, J. Y. J. 2010. Context-aware taxi demand hotspots prediction. *Int. J. Bus. Intell. Data Min. 5*, 1, 3–18.

Chang, H.-W., Tai Y.-C., Chen, H. W., and Hsu, J. Y.-J. 2008. iTaxi: Context-aware taxi demand hotspots prediction using ontology and data mining approaches. In *Proceedings of the 13th Conference on Artificial Intelligence and Applications (TAAI'08)*.

Chawathe, S. S. 2007. Segment-based Map Matching. In *Proceedings of the IEEE Intelligent Vehicles Symposium*. 1190–1197.

Chawla, S., Zheng, Y., and Hu, J. 2012. Inferring the root cause in road traffic anomalies. In *Proceedings of the IEEE 12th International Conference on Data Mining (ICDM'12)*. 141–150.

Chen, C., Zhang, D., Castro, P. S., Li, N., Sun, L., and Li, S. 2011. Real-time detection of anomalous taxi trajectories from GPS traces. In *Proceedings of the International ICST Conference on Mobile and Ubiquitous Systems*. 63–74.

Chen, C., Zhang, D., Zhou, Z.-H., Li, N., Atmaca, T., and Li, S. 2013. B-Planner: Night bus route planning using large-scale taxi GPS traces. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom'13)*.

Chen, G., Chen, B., and Yu, Y. 2010a. Mining frequent trajectory patterns from GPS tracks. In *Proceedings of the International Conference on Computational Intelligence and Software Engineering (CiSE'10)*. 1–6.

Chen, G., Jin, X., and Yang, J. 2010b. Study on spatial and temporal mobility pattern of urban taxi services. In *Proceedings of the International Conference on Intelligent Systems and Knowledge Engineering (ISKE'10)*. 422–425.

Chen, P.-Y. 2010. A fuel-saving and pollution-reducing dynamic taxi-sharing protocol in VANETs. In *Proceedings of the IEEE 72nd Vehicular Technology Conference Fall (VTC 2010-Fall)*. 1–5.

Chen, Y. and Krumm, J. 2010. Probabilistic modeling of traffic lanes from GPS traces. In *Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 81–88.

Cooke, K. L. and Halsey, E. 1966. The shortest route through a network with time-dependent internodal transit times. *J. Math. Anal. Appl. 14*, 493–498.

Crandall, D. J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., and Kleinberg, J. 2010. Inferring social ties from geographic coincidences. *PNAS 107*, 52, 22436–22441.

Culotta, A. 2010. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the Workshop on Social Media Analytics*. 115–122.

Demirbas, M., Bayir, M. A., Akcora, C. G., Yilmaz, Y. S., and Ferhatosmanoglu, H. 2010. Crowd-sourced sensing and collaboration using Twitter. In *Proceedings of the IEEE International Symposium on a World of Wireless, Mobile, and Multimedia Networks*. 1–9.

Ding, B., Yu, J. X., and Qin, L. 2008. Finding time-dependent shortest paths over large graphs. In *Proceedings of the International Conference on Extending Database Technology: Advances in Database Technology*. 205–216.

d'Orey, P. M. 2012. Empirical evaluation of a dynamic and distributed taxi-sharing system. In *Proceedings of the 15th International IEEE Conference on Intelligent Transportation Systems (ITSC'12)*. 140–146.

Duda, R. O., Hart, P. E., and Stork, D. G. 2001. *Pattern Classification*. John Wily & Sons, New York.

Eagle, N. and Pentland, A. 2006. Reality mining: Sensing complex social systems. *Pers. Ubiquit. Comput. 10*, 4, 255–268.

Eagle, N., Pentland, A., and Lazis, D. 2009. Inferring friendship network structure by using mobile phone data. *PNAS 106*, 36, 15274–15278.

Edelkamp, S. and Schrödl, S. 2003. *Route Planning and Map Inference with Global Positioning Traces*. Springer-Verlag, New York, NY, 128–151.

Farrahi, K. and Gatica-Perez, D. 2008. What did you do today?: Discovering daily routines from large-scale mobile data. In *Proceedings of the ACM International Conference on Multimedia*. 849–852.

Farrahi, K. and Gatica-Perez, D. 2011. Discovering routines from large-scale human locations using probabilistic topic models. *ACM TIST 2*, 1, 3:1–3:27.

Fathi, A. and Krumm, J. 2010. Detecting road intersections from GPS traces. In *Proceedings of the International Conference on Geographic Information Science*. 56–69.

Finkel, R. and Bentley, J. L. 1974. Quad trees: A data structure for retrieval on composite keys. *Acta Inf. 4*, 1, 1–9.

Froehlich, J. and Krumm, J. 2008. Route prediction from trip observations. In *Proceedings of the Intelligent Vehicle Initiative (IVI) Technology Advanced Controls and Navigation Systems, SAE World Congress and Exhibition*.

Froehlich, J., Neumann, J., and Oliver, N. 2009. Sensing and predicting the pulse of the city through shared bicycling. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 1420–1426.

Fuchs, H., Kedem, Z. M., and Naylor, B. F. 1980. On visible surface generation by a priori tree structures. In *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques*. 124–133.

Ge, Y., Xiong, H., Liu, C., and Zhou, Z.-H. 2011. A taxi driving fraud detection system. In *Proceedings of the IEEE International Conference on Data Mining*. 181–190.

Ge, Y., Xiong, H., Tuzhilin, A., Xiao, K., Gruteser, M., and Pazzani, M. J. 2010. An energy-efficient mobile recommender system. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 899–908.

Giannotti, F., Nanni, M., Pedreschi, D., and Pinelli, F. 2007. Trajectory pattern mining. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 330–339.

Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S., and Trasarti, R. 2011. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *VLDB J. 20*, 5, 695–719.

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

Gonzalez, H., Han, J., Li, X., Myslinska, M., and Sondag, J. P. 2007. Adaptive fastest path computation on a road network: A traffic mining approach. In *Proceedings of the International Conference on Very Large Data Bases*. 794–805.

González, M. C., Hidalgo, C. A., and Barabási, A.-L. 2008. Understanding individual human mobility patterns. *Nature 453*, 779–782.

Greenfeld, J. 2002. Matching GPS observations to locations on a digital map. In *Proceedings of the 81st Annual Meeting of the Transportation Research Board*.

Gühnemann, A., Schäfer, R.-P., Thiessenhusen, K.-U., and Wagner, P. 2004. Monitoring traffic and emissions by floating car data. Working paper ITS-WP-04-07. Institute of Transportation Studies, University of Sydney, Australia.

Guttman, A. 1984. R-trees: A dynamic index structure for spatial searching. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 47–57.

Haklay (Muki), M. and Weber, P. 2008. OpenStreetMap: User-generated street maps. *IEEE Pervasive Comput. 7*, 4, 12–18.

Herring, R., Hofleitner, A., Abbeel, P., and Bayen, A. 2010. Estimating arterial traffic conditions using sparse probe data. In *Proceedings of the International IEEE Conference on Intelligent Transportation Systems*. 929–936.

Hu, H., Wu, Z., Mao, B., Zhuang, Y., Cao, J., and Pan, J. 2012b. Pick-up tree based route recommendation. In *Proceedings of the International Conference on Web-Age Information Management*. 471–483.

Hu, J., Cao, W., Luo, J., and Yu, X. 2009. Dynamic modeling of urban population travel behavior based on data fusion of mobile phone positioning Ddata and FCD. In *Proceedings of the International Conference on Geoinformatics*. 1–5.

Hu, X., Gao, S., Chiu, Y.-C., and Lin, D.-Y. 2012a. Modeling routing behavior for vacant taxi cabs in urban traffic networks. *Transport. Res. Rec. 2284*, 81–88.

Huang, H., Zhu, Y., Li, X., Li, M., and Wu, M.-Y. 2010. META: A mobility model of MEtropolitan TAxis extracted from GPS traces. In *Proceedings of the Wireless Communications and Networking Conference*. 1–6.

Jiang, B., Yin, J., and Zhao, S. 2009. Characterizing the human mobility pattern in a large street network. *Phys. Rev. E 80*, 021136.

Jolliffe, I. T. 1986. *Principal Component Analysis*. Springer-Verlag.

Kanoulas, E., Du, Y., Xia, T., and Zhang, D. 2006. Finding fastest paths on a road network with speed patterns. In *Proceedings of the International Conference on Data Engineering*. 10.

Krumm, J. and Horvitz, E. 2006. Predestination: Inferring destinations from partial trajectories. In *Proceedings of the International Conference on Ubiquitous Computing*. 243–260.

Lee, J., Shin, I., and Park, G.-L. 2008. Analysis of the passenger pick-up pattern for taxi location recommendation. In *Proceedings of the International Conference on Networked Computing and Advanced Information Management*. 199–204.

Li, B., Zhang, D., Sun, L., Chen, C., Li, S., Qi, G., and Yang, Q. 2011b. Hunting or waiting? Discovering passenger-finding strategies from a large-scale real-world taxi dataset. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*. 63–68.

Li, N. and Chen, G. 2009. Multi-layered friendship modeling for location-based mobile social networks. In *Proceedings of the International Conference on Mobile and Ubiquitous Systems: Computing Networking and Services*. 1–10.

Li, Q., Zheng, Z., Yang, B., and Zhang, T. 2009a. Hierarchical route planning based on taxi GPS-trajectories. In *Proceedings of the International Conference on Geoinformatics*. 1–5.

Li, Q., Zheng, Z., Zhang, T., Li, J., and Wu, Z. 2011c. Path-finding through flexible hierarchical road networks: An experiential approach using taxi trajectory data. *Int. J. Appl. Earth Obs. 13*, 1, 110–119.

Li, X., Li, M., Shu, W., and Wu, M. 2007. A practical map-matching algorithm for GPS-based vehicular networks in Shanghai urban area. In *Proceedings of the IET Conference on Wireless, Mobile, and Sensor Networks*. 454–457.

Li, X., Li, Z., Han, J., and Lee, J.-G. 2009b. Temporal outlier detection in vehicle traffic data. In *Proceedings of the International Conference on Data Engineering*. 1319–1322.

Li, X., Pan, G., Qi, G., and Li, S. 2011a. Predicting urban human mobility using large-scale taxi traces. In *Proceedings of the First Workshop on Pervasive Urban Applications*.

Liao, L., Fox, D., and Kautz, H. 2007. Learning and inferring transportation routines. *Artif. Intell. 171*, 5–6, 311–331.

Liao, Z. 2003. Real-time taxi dispatching using global positioning systems. *Commun. ACM 46*, 5, 81–83.

Liao, Z., Yu, Y., and Chen, B. 2010. Anomaly detection in GPS data based on visual analytics. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*. 51–58.

Lim, S., Balakrishnan, H., Gifford, D., Madden, S., and Rus, D. 2010. Stochastic motion planning and application to traffic. *Int. J. Robot. Res. 30*, 699–712.

Lin, Y., Li, W., Qiu, F., and Xu, H. 2012. Research on optimization of vehicle routing problem for ride-sharing taxi. In *Proceedings of the 8th International Conference on Traffic and Transportation Studies (ICTTS'12)*.

Lippi, M., Bertini, M., and Frasconi, P. 2010. Collective traffic forecasting. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II*. 259–273.

Liu, L., Andris, C., Biderman, A., and Ratti, C. 2009a. Uncovering taxi driver's mobility intelligence through his trace. *IEEE Pervasive Comput*.

Liu, L., Andris, C., and Ratti, C. 2010a. Uncovering cabdrivers' behavior patterns from their digital traces. *Comput. Environ. Urban Syst. 34*, 6, 541–548.

Liu, L., Biderman, A., and Ratti, C. 2009b. Urban mobility landscape: Real time monitoring of urban mobility patterns. In *Proceedings of the International Conference on Computers in Urban Planning and Urban Management*.

Liu, S., Liu, Y., Ni, L. M., Fan, J., and Li, M. 2010b. Towards mobility-based clustering. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 919–928.

Liu, W., Zheng, Y., Chawla, S., Yuan, J., and Xie, X. 2011. Discovering spatio-temporal causal interactions in traffic data streams. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1010–1018.

Liu, X., Zhu, Y., Wang, Y., Forman, G., Ni, L. M., Fang, Y., and Li, M. 2012a. Road recognition using coarse-grained vehicular traces. Tech. rep. HPL-2012-26. HP Laboratories.

Liu, Y., Kang, C., Gao, S., and Xiao, Y. 2012b. Understanding intra-urban trip patterns from taxi trajectory data. *J. Geogr. Syst. 14*, 4, 463–483.

Liu, Y., Wang, F., Xiao, Y., and Gao, S. 2012c. Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai. *Landscape Urban Plan. 106*, 1, 73–87.

Lou, Y., Zhang, C., Zheng, Y., Xie, X., Wang, W., and Huang, Y. 2009. Map-matching for low-sampling-rate GPS trajectories. In *Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 352–361.

Ma, S., Zheng, Y., and Wolfson, O. 2013. T-Share: A large-scale dynamic taxi ridesharing service. In *Proceedings of the IEEE Conference on Data Engineering (ICDE'13)*.

Martino, M., Britter, R., Outram, C., Zacharias, C., Biderman, A., and Ratti, C. 2010. Senseable City. In *Digital Urban Modelling and Simulation*.

Miluzzo, E., Lane, N. D., Fodor, K., Peterson, R., Lu, H., Musolesi, M., Eisenman, S. B., Zheng, X., and Campbell, A. T. 2008. Sensing meets mobile social networks: The design, implementation and evaluation of the CenceMe application. In *Proceedings of the ACM Conference on Embedded Networked Sensor Systems*. 337–350.

Monreale, A., Pinelli, F., Trasarti, R., and Giannotti, F. 2009. WhereNext: A location predictor on trajectory pattern mining. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 637–646.

Ng, A. Y. and Russell, S. 2000. Algorithms for inverse reinforcement learning. In *Proceedings of the International Conference on Machine Learning*. 663–670.

Palma, A. T., Bogorny, V., Kuijpers, B., and Alvares, L. O. 2008. A clustering-based approach for discovering interesting places in trajectories. In *Proceedings of the ACM Symposium on Applied Computing*. 863–868.

Pang, L. X., Chawla, S., Liu, W., and Zheng, Y. 2011. On mining anomalous patterns in road traffic streams. In *Proceedings of the International Conference on Advanced Data Mining and Applications: Volume Part II*. 237–251.

Patterson, D. J., Liao, L., Fox, D., and Kautz, H. 2003. Inferring high-level behavior from low-level sensors. In *Proceedings of the International Conference on Ubiquitous Computing*. 73–89.

Peng, C., Jin, X., Wong, K.-C., Shi, M., and Liò, P. 2012. Collective human mobility patter from taxi trips in urban area. *PLoS ONE 7*, 4, e34487.

Phithakkitnukoon, S., Veloso, M., Biderman, A., Bento, C., and Ratti, C. 2010. Taxi-aware map: Identifying and predicting vacant taxis in the city. In *Proceedings of the International Conference on Ambient Intelligence*. 86–95.

Powell, J. W., Huang, Y., Bastani, F., and Ji, M. 2011. Towards reducing taxicab cruising time using spatio-temporal profitability maps. In *Proceedings of the International Conference on Advances in Spatial and Temporal Databases*. 242–260.

Puterman, M. L. 1994. *Markov Decision Processes*. John Wily & Sons, New York, NY.

Qi, G., Li, X., Li, S., Pan, G., and Wang, Z. 2011. Measuring social functions of city regions from large-scale taxi behaviors. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*. 384–388.

Qian, S., Zhu, Y., and Li, M. 2012. Smart recommendation by mining large-scale GPS traces. In *Proceedings of the Wireless Communications and Networking Conference*. 3267–3272.

Ratti, C., Pulselli, R. M., Williams, S., and Frenchman, D. 2009. Mobile landscapes: Using location data from cell phones for urban analysis. *Environ. Plann. B 33*, 5, 727–748.

Rogers, S., Langley, P., and Wilson, C. 1999. Mining GPS data to augment road models. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 104–113.

Ross, S. 2006. *Simulation*. Academic Press.

Sakoe, H. and Chiba, S. 1990. Dynamic programming algorithm optimization for spoken word recognition. In *Readings in Speech Recognition*, A. Waibel and K.-F. Lee, Eds., Morgan Kaufmann Publishers, 159–165.

Sanders, P. and Schultes, D. 2005. Highway hierarchies hasten exact shortest path queries. In *Proceedings of the European Conference on Algorithms*. 568–579.

Schäfer, R.-P., Thiessenhusen, K.-U., and Wagner, P. 2002. A traffic information system by means of real-time floating-car data. In *Proceedings of the World Congress on Intelligent Transport Systems*.

Scholkopf, B. and Smola, A. J. 2002. *Learning with Kernels*. MIT Press.

Schroedl, S., Wagstaff, K., Rogers, S., Langley, P., and Wilson, C. 2004. Mining GPS traces for map refinement. *Data Min. Knowl. Disc. 9*, 59–87.

Song, C., Koren, T., Wang, P., and Barabási, A.-L. 2010a. Modelling the scaling properties of human mobility. *Nature Phys. 6*, 818–823.

Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. 2010b. Limits of predictability in human mobility. *Science 327*, 5968, 1018–1021.

Su, H. and Yu, S. 2007. Hybrid GA based online support vector machine model for short-term traffic flow forecasting. In *Proceedings of the International Conference on Advanced Parallel Processing Technologies*. 743–752.

Sun, L., Zhang, D., Chen, C., Castro, P. S., Li, S., and Wang, Z. 2012. Real time anomalous trajectory detection and analysis. In *Mobile Netw. Appl. 18*, 3, 341–356.

Takayama, T., Matsumoto, K., Kumagai, A., Sato, N., and Murata, Y. 2011. Waiting/cruising location recommendation for efficient taxi business. *Int. J. Syst. Appl. Eng. Dev. 5*, 2, 224–236.

Tao, C.-C. 2007. Dynamic taxi-sharing service using intelligent transportation system technologies. In *Proceedings of the International Conference on Wireless Communications, Networking, and Mobile Computing (WiCom'07)*.

Veloso, M., Phithakkitnukoon, S., and Bento, C. 2011a. Sensing urban mobility with taxi flow. In *Proceedings of the ACM SIGSPATIAL International Workshop on Location-Based Social Networks*. 41–44.

Veloso, M., Phithakkitnukoon, S., and Bento, C. 2011b. Urban mobility study using taxi traces. In *Proceedings of the International Workshop on Trajectory Data Mining and Analysis*. 23–30.

Veloso, M., Phithakkitnukoon, S., Bento, C., Fonseca, N., and Olivier, P. 2011c. Exploratory study of urban flow using taxi traces. In *Proceedings of the 1st Workshop on Pervasive Urban Applications*.

Šingliar, T. and Hauskrecht, M. 2007. Modeling highway traffic volumes. In *Proceedings of the European Conference on Machine Learning*. 732–739.

Šingliar, T. and Hauskrecht, M. 2008. Approximation strategies for routing in dynamic stochastic networks. In *Proceedings of the International Symposium on Artificial Intelligence and Mathematics*.

Wang, H., Zou, H., Yue, Y., and Li, Q. 2009. Visualizing hot spot analysis result based on Mashup. In *Proceedings of the International Workshop on Location Based Social Networks*. 45–48.

Wang, Y., Zhu, Y., He, Z., Yue, Y., and Li, Q. 2011. Challenges and opportunities in exploiting large-scale GPS probe data. Tech. rep. HPL-2011-109, HP Laboratories.

Wen, H., Hu, Z., Guo, J., Zhu, L., and Sun, J. 2008. Operational analysis on Beijing Road network during the Olympic Games. *J. Trans. Sys. Eng. Info. Tech. 8*, 6, 32–37.

Worrall, S. and Nebot, E. 2007. Automated process for generating digitised maps through GPS data compression. In *Proceedings of the Australasian Conference on Robotics and Automation*.

Xu, R. and Wunsch, D. 2005. Survey of clustering algorithms. *IEEE Trans. Neural Netw. 16*, 645–678.

Xue, A. Y., Zhang, R., Zheng, Y., Xie, X., Huang, J., and Xu, Z. 2013. Destination prediction by sub-trajectory synthesis and privacy protection against such prediction. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE'13)*.

Yamamoto, K., Uesugi, K., and Watanabe, T. 2010. Adaptive routing of cruising taxis by mutual exchange of pathways. In *Proceedings of the International Conference on Knowledge-Based Intelligent Information and Engineering Systems: Part II*. 559–566.

Yin, H. and Wolfson, O. 2004. A weight-based map matching method in moving objects databases. In *Proceedings of the International Conference on Scientific and Statistical Database Management*. 437–438.

Yuan, J. and Zheng, Y. 2010. T-Drive: driving directions based on taxi trajectories. In *Proceedings of the SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 99–108.

Yuan, J., Zheng, Y., Xie, X., and Sun, G. 2011a. Driving with knowledge from the physical world. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 316–324.

Yuan, J., Zheng, Y., Xie, X., and Sun, G. 2013. T-Drive: Enhancing driving directions with taxi drivers' intelligence. *IEEE Trans. Knowl. Data Eng. 25*, 1, 220–232.

Yuan, J., Zheng, Y., Zhang, C., Xie, X., and Sun, G. 2010. An interactive voting-based map matching algorithm. In *Proceedings of the International Conference on Mobile Data Management*. 43–52.

Yuan, J., Zheng, Y., Zhang, L., Xie, X., and Sun, G. 2011b. Where to find my next passenger? In *Proceedings of the International Conference on Ubiquitous Computing*. 109–118.

Yuan, J., Zheng, Y., and Xie, X. 2012a. Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 186–194.

Yuan, N. J., Zheng, Y., and Xie, X. 2012b. Segmentation of urban areas using road networks. Tech. rep. MSR-TR-2012-65.

Yuan, N. J., Zheng, Y., Zhang, L., and Xie, X. 2011c. T-Finder: A recommender system for finding passengers and vacant taxis. *IEEE Trans. Knowl. Data Eng.*

Yue, Y., Zhuang, Y., Li, Q., and Mao, Q. 2009. Mining time-dependent attractive areas and movement patterns from taxi trajectory data. In *Proceedings of the International Conference on Geoinformatics*. 1–6.

Zhang, D., Guo, B., and Yu, Z. 2011a. The emergence of social and community intelligence. *Computer 44*, 7, 21–28.

Zhang, D., Li, N., Zhou, Z.-H., Chen, C., Sun, L., and Li, S. 2011b. iBAT: Detecting anomalous taxi trajectories from GPS traces. In *Proceedings of the International Conference on Ubiquitous Computing*. 99–108.

Zhang, W., Li, S., and Pan, G. 2012. Mining the semantics of origin-destination flows using taxi traces. In *Proceedings of the Workshop of Ubiquitous Computing*. 943–949.

Zhang, W., Xu, J., and Wang, H. 2007. Urban traffic situation calculation methods based on probe vehicle data. *J. Transport. Syst. Eng. Inform. Technol. 7*, 1, 43–49.

Zheng, X., Liang, X., and Xu, K. 2012. Where to wait for a taxi? In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*. 149–156.

ZHENG, Y., LIU, L., WANG, L., AND XIE, X. 2008. Learning transportation mode from raw GPS data for geographic applications on the Web. In *Proceedings of the International Conference on World Wide Web*. 247–256.

ZHENG, Y., LIU, Y., YUAN, J., AND XIE, X. 2011. Urban computing with taxicabs. In *Proceedings of the International Conference on Ubiquitous Computing*. 89–98.

ZHENG, Y., XIE, X., AND MA, W.-Y. 2010. GeoLife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull. 33*, 2, 32–39.

ZHENG, Y., ZHANG, L., XIE, X., AND MA, W.-Y. 2009. Mining interesting locations and travel sequences from GPS trajectories. In *Proceedings of the International Conference on World Wide Web*. 791–800.

ZIEBART, B. D., MAAS, A. L., DEY, A. K., AND BAGNELL, J. A. 2008. Navigate like a cabbie: Probabilistic reasoning from observed context-aware behavior. In *Proceedings of the International Conference on Ubiquitous Computing*. 322–331.

Received April 2012; revised November 2012; accepted February 2013