Using Trajectories of Moving Objects in Traffic Prediction and Management

Győző Gidófalvi, Ehsan Saqib

The Royal Institute of Technology (KTH), Geoinformatics, Drottning Kristinas väg 30, 100 44 Stockholm, Sweden
Email: gyozo.gidofalvi@abe.kth.se, esaqib@kth.se

1. Introduction

The rapid growth of demand for transportation, high levels of car dependency caused by the urban sprawl have exceeded the slow increments in transportation infrastructure supply in many areas causing severe traffic congestion. Some well-known negative effects of traffic congestion include: the fuel wasted in idling vehicles leading to increasing air pollution and carbon dioxide emissions; the time and associated cost lost by motorists sitting in traffic jams; the wear-and-tear on vehicles and infrastructure as a result of stop-and-go traffic; and, last but not least, the inflicted stress and fatigue on motorist causing unnecessary accidents.

In dense urban areas, adding capacity through construction of new facilities is difficult due to lack of space and prohibitive costs. A more viable approach to cope with the congestion problem is to monitor traffic congestion, understand the causes of its formation and development, and use the aforementioned knowledge in traffic management systems and transportation planning to mitigate traffic congestion.

Earlier efforts to derive information about traffic congestion have used fixed location sensors. Portals have been built in highways for collecting information such as flow and punctual speed at certain locations. Such data collection methods require huge building costs and provide data that allows analysis methods to gain limited knowledge about the causes of traffic congestion formation and development. Currently, traffic monitoring centers consider deriving information about traffic congestion, in particular traffic parameters such as flow and travel time, using both fixed and mobile data collection methods. Studies using empirical data (Morán C and Bang KL, 2010) have analyzed data provided by two methods for estimating travel-time: Automatic Travel Time System – ATTS (that identifies number plates at intersections and later matches them) and, Floating Car Surveys – FCS (where specialized vehicles cover a route back and forth during the survey period). It was observed that ATTS provide in large part inaccurate or unfeasible data and FCS have a low sample rate and high cost (Morán C and Bang KL, 2010).

More recently, as proposed by Gidófalvi and Morán (2010) and many others, the increasing availability and accuracy of positioning technologies (primarily GPS) embedded in on-board navigation systems of both private and commercial vehicles and mobile devices that are carried by the drivers enable the low-cost FCS-like data collection from large numbers of private vehicles. However, such data collection can poses serious privacy threats as the trajectories collected refer to the highly sensitive movements of private individuals. To this extent, the first part (Section 2) of this short paper outlines an approach that using trajectories of moving objects in a privacy-preserving manner, in real-time, monitors traffic congestion, extracts knowledge about the causes of congestion formation and development, and uses the extracted knowledge in a traffic prediction and management tasks to mitigate congestion. To aid the development and assessment of traffic prediction and management systems that use moving object trajectories, the second part of this short paper (Section 4) proposes the development-, and highlights important aspects of the design of a benchmark for traffic prediction and management based on a specific data set.

2. Movement Pattern Based Traffic Prediction and Management

The proposed system has a client-server architecture in which location-aware mobile clients map match (Jensen and Tradisauskas 2009) their locations to road segments, perform road-network based location anonymization, and report their trajectories to the server in form of a continuous stream of time-stamped traversed road segments. The server stores the evolving route trajectories of clients in a compressed format using *route-tree*, which due to length limitations is not described here. Simultaneously, using a sliding window model in an incremental and continuous fashion (Mozafari et al. 2008, Jiang and Gruenwald 2006) the server extracts and stores movement and traffic patterns in form of frequent (sub-)routes and congested road segments (Gidófalvi and Pedersen 2009). The server uses the extracted knowledge for traffic and congestion prediction at a given time point by combining the information about the partial-routes in the route-tree with relevant historical frequent routes to estimate the expected number and speed of clients on each road segment in the near future. Finally, based on the estimates the server performs traffic management by providing various traffic information / advisory services (estimated travel times, variable speed advisory, alternative routes) relating to actual and predicted traffic conditions / events (accidents, road construction and congestion) to the likely-to-be-affected clients / vehicles. The following subsections further elaborate on important aspects of the proposed system.

2.1 Location Privacy and Anonymization

Trajectories of individuals contain highly sensitive personal information; therefore, to protect the privacy of individuals, they need to be adequately anonymized. A major privacy threat is the identification of individuals by *self-correlating* trajectories to identify frequently visited private locations, e.g., home, work and subsequently *cross-referencing* a subset of these private locations to publically available external data sources, e.g., Yellow pages (Gidófalvi et al. 2010).

A number of privacy protection frameworks have been proposed to protect against the above described threat. A common approach is to *generalize* the exact locations of individuals to *cloaking region*. Following the traditional notion of *k-anonymity*, most privacy protection frameworks construct cloaking regions such that at the time of the location report there are at least *k* objects in the given region. As identified in (Gidófalvi et al. 2010), there are a number of problems with this method. First, determining the *k-anonymity* based cloaking region requires trusted components, e.g., a server, often termed as the *anonymizer*, that is aware of the exact positions of the objects. Second, as the cloaking regions depend on the positions of nearby objects, for a given location the cloaking region varies over time depending on the density of the objects, allowing an attacker to infer the locations of the private location to be within the intersection of the reported cloaking regions for the location. Finally, *k-anonymity* based cloaking regions are likely to be over-protective in low density rural areas, and under-protective in high density, sensitive hot spots, e.g., red light district.

To overcome the above shortcomings, based on prior work by Gidófalvi et al. 2010, the proposed system adopts an privacy protection framework in which clients specify their requirements of location privacy, based on the notions of *anonymization road segment sets* and *location probabilities*, intuitively saying how precisely they want to be located in given areas. Such a privacy protection framework serves the transportation application domain well for two reasons. First, it allows clients to set their privacy requirements to arbitrarily low values in areas where the threat of being identified is very low, which arguably constitutes most parts of the routes – only excluding the parts very near the origin and destination of the route. Second, road-network based generalization of noisy location readings allow aggregation based data mining methods the extraction of accurate and relevant movement patterns in the transportation application domain (Gidófalvi and Pedersen 2009).

2.2 Movement and Traffic Patterns

In recent past, a large number of methods have been proposed to extract movement and traffic patterns. Dodge et al. (2008) provide a good review and taxonomy of these methods. For the target application at hand the most promising patterns include trajectory clusters (Rinzivillo et al. 2008) and frequent spatio-temporal sequences, i.e., routes (Gidófalvi and Pedersen 2009). The proposed system adopts the latter for two reasons. First, the trajectory representation through road-network based generalization allows well-researched data mining methods, such as (maximal/closed) frequent itemset / sequential pattern mining methods to efficiently extract frequent routes. Second, the so extracted frequent routes can be efficiently stored- and their relationships to each other can be easily and efficiently queried in a DBMS.

To preserve clarity, Figure 1 shows the two dimensional projection of such frequent routes. Figure 1 clearly shows that each segment of every pattern has two attributes: a *vehicle count* and a *speed*. What Figure 1 fails to illustrate is that such patterns also have a *direction* and a *spatial and temporal relationship between each other*. All four of these pattern attributes are vital for accurate traffic prediction and provide insight into the creation and development of congestions.

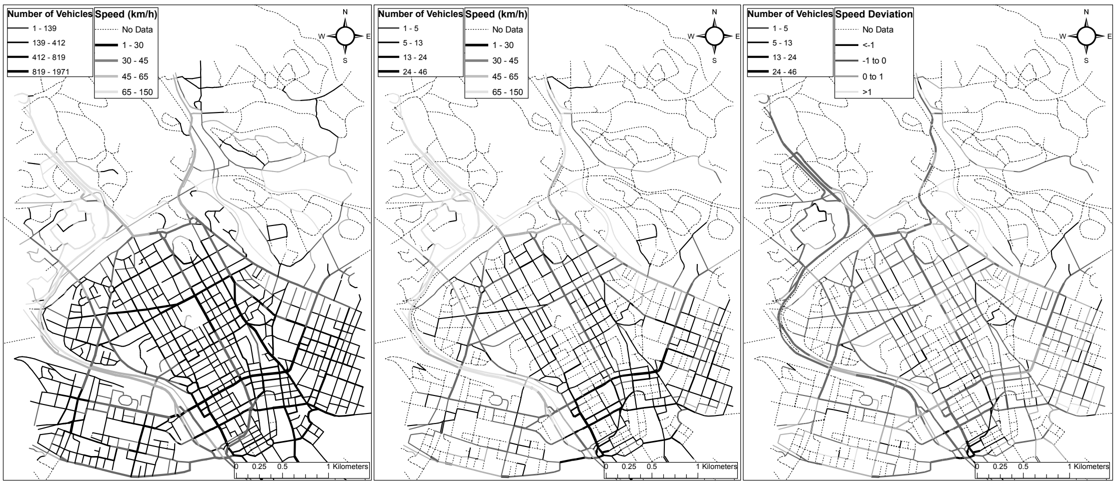

Figure 1: Frequent routes with speed profiles. Daily frequent routes and speed profiles (left). 8am frequent routes and speed profiles (center). 8am speed deviations (right).

3. Empirical Study Based on Real World Trajectories

The performance and scalability of the proposed system will be tested on the following real world trajectory data set (stream). The trajectory data set is provided by Trafik Stockholm and is available at the Transport and Logistic Division of the Department of Urban Planning and Environment, Royal Institute of Technology (KTH), Sweden (collaborating partner). The trajectory data set contains the GPS readings of 1500 taxis and 400 trucks travelling on the streets of Stockholm. Each taxi produces a reading once every 60 seconds approximately. This reading includes only taxi identification and location information. Taxis produce readings less frequently when they are not carrying any passengers. Trucks use more recent and more accurate GPS devices that produce readings once every 30 seconds and include identification, location, speed and heading information. The peak data rate for the whole city is over 1000 readings per minute, and there are approximately 170 million readings during the course of a year.

The study will assess the performance of the system in terms of prediction accuracy and throughput and will evaluate demonstrate the scalability of the approach by replaying the data at several times the actual data rate. The implementation of the system will utilize an IBM InfoSphere Streams parallel and distributed DSMS running on a cluster of commodity hardware.

4. Development of a Benchmark

Several trends in urban mobility put the development of effective traffic prediction and management systems in high demand. Early systems have primarily used punctuated speed and flow measurements from fixed location sensors in conjunction with traffic models to tackle the prediction and management tasks. More recently (as in the current proposal), fuelled by the wide-spread adoption of GPS-based on-board navigation systems and location-aware mobile devices, to improve accuracy the use of moving object trajectories in such systems have been proposed. In such proposals (as in the present proposal) it is assumed that vehicles periodically submit their location to a central server. In turn the server extracts mobility / traffic patterns from the submitted locations. The extracted mobility / traffic patterns, together with the current locations of the vehicles, are both used in short- and long term traffic prediction, management and planning tasks. Additionally, the current and near-future traffic conditions are sent in real-time to the likely-to-be-affected vehicles. To aid the development and assessment of such systems, a benchmark for traffic prediction and management is needed.

4.1 Benchmark Data Set

One possible real world moving object trajectory data set (stream) for developing such a traffic prediction and management benchmark is described in Section 3. Two notable characteristics of the described trajectory data set are that it represents a relatively large sample of the moving object population, but locations of samples are obtained relatively infrequently. The latter characteristics is effectively visualized in Section 4.3

4.2 Benchmark Tests

Benchmark tests are to be developed with clearly defined criteria for evaluating and comparing the *performance* and *accuracy* of systems for the traffic prediction and management tasks outlined in Section 2. Benchmark test that evaluate the *scalability* of systems should also be designed. The development of benchmark tests, given the data and the application scenario, should carefully examine and address the following issues:

- **Trajectory sample bias**: Taxi trajectories have a different spatial and temporal distribution from that of private vehicle trajectories.
- **Absence of individual mobility patterns**: Taxi trajectories do not contain *periodic* patterns that pertain to the mobility of individual persons. Consequently, benchmark tests developed based on the taxi trajectories cannot be used to adequately assess the performance of systems that exploit such patterns.
- **Need for privacy**: Privacy is a major issue in the application setting, therefore benchmark tests should be able to assess the performance and accuracy of systems

for different privacy levels, which are provided by some location anonymization framework for example by blurring/cloaking locations of objects.

- **Realistic scalability tests**: Replaying data at several times the actual data rate can only partially simulate higher input data rates. However, given this data scaling method, the accurate evaluation of scalability is questionable, because the number of patterns in the data does not increase, i.e., the processing time for pattern extraction and management remains constant or only increases linearly with the scaling factor.

4.3 Visualization of Raw Trajectory Data and Derived Mobility Patterns

To gain a better understanding of the nature of the proposed trajectory data to be used in the development of the proposed benchmark, Figure 2 visualizes a *small subset* of the raw data and the extracted mobility patterns. In particular, the subset is spatially restricted to the inner city of Stockholm; the area of the full spatial extent of the complete trajectory data is approximately 25 times larger than the area of the extent that is visualized in Figure 2. Furthermore, the subset is also temporally restricted and only includes data for the course of a single day. Given this subset, the data visualized in Figure 2(left) is further restricted to only contain the raw GPS readings of 100 taxis. Subsequently, the data visualized in Figure 2(center) is even further restricted to only contain the raw 2D trajectories of 10 taxis. Figure 2(center) clearly illustrates one consequence of the low sampling rate and visualizes the *linear interpolation* of the trajectories. Finally, Figure 2(right) shows the extracted mobility patterns, i.e., frequent routes with speed profiles, which, given the previously described spatio-temporal extent, are extracted from the raw data of *all* 1500 taxis using *road network based interpolation* of the trajectories and frequent pattern mining techniques.

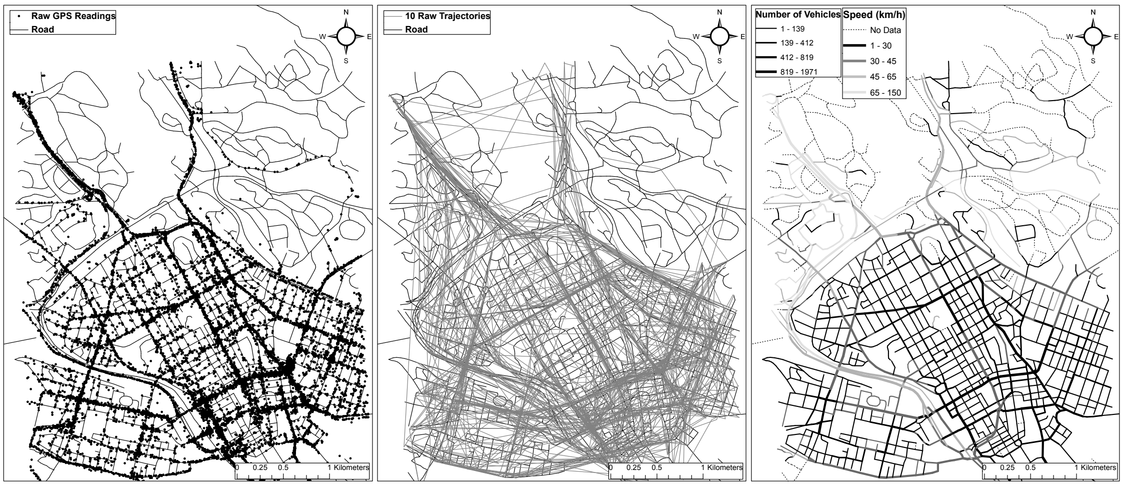

Figure 2: Raw Trajectory Data and Derived Traffic Patterns. Raw GPS readings of 100 taxis (left). Raw trajectories of 10 taxis (center). Frequent routes of all taxis (right).

5. Conclusions

Motivated by the accelerating need for effective traffic prediction and management systems and the new possibilities for such systems allowed by GPS-enabled mobile devices, the paper outlined an approach for the privacy-preserving usage of moving object trajectories in a traffic prediction and management system. Furthermore, to aid the development and assessment of traffic prediction and management systems that use moving object trajectories, the paper also proposed the development-, and highlighted important aspects of the design of a benchmark for traffic prediction and management based on a specific data set.

References

Dodge S, Weibel R, and Lautenschütz A-K, 2008, Towards a Taxonomy of Movement Patterns. *Information Visualization*, 7(3):240–252.

Jensen CS and Tradisauskas N, 2009, Map Matching. *Encyclopedia of Database Systems 2009*:1692–1696.

Jiang N and Gruenwald L, 2006, CFI-Stream: Mining Closed Frequent Itemsets in Data Streams. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, USA, 592–597.

Gidófalvi G and Pedersen TB, 2009, Mining Long, Sharable Patterns in Trajectories of Moving Objects. *Geoinformatica*, 13(1):27–55.

Gidófalvi G, Huang X, and Pedersen TB, 2010. Probabilistic Grid-Based Approaches for Privacy Preserving Data Mining on Moving Object Trajectories. In: Bonchi F, Ferrari E (eds), *Privacy-Aware Knowledge Discovery: Novel Applications and New Techniques*. CRC PRESS.

Gidófalvi G and Morán C, 2010, Estimating Traffic Performance in Road Networks from Anonymized GPS Vehicle Probes. *Workshop on Movement Research: Are you in the flow? at the 13th AGILE International Conference on Geographic Information Science*.

Morán C and Bang KL, 2010, Reliability of Congestion Performance Measures. In: *Proceedings of the Institution of Civil Engineers - Transport*.

Mozafari B, Thakkar H, and Zaniolo C, 2008, Verifying and Mining Frequent Patterns from Large Windows over Data Streams. In: *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, Cancun, Mexico, 179–188.

Rinzivillo S, Pedreschi D, Nanni M, Giannotti F, Andrienko N and Andrienko G, 2008, Visually Driven Analysis of Movement Data by Progressive Clustering. *Information Visualization*, 7(3):225–239.